

Robust Linear Regression Analysis - A Greedy Approach

George Papageorgiou*, Pantelis Bouboulis[†] and Sergios Theodoridis[‡]

May 11, 2015

Abstract

The task of robust linear estimation in the presence of outliers is of particular importance in signal processing, statistics and machine learning. Although the problem has been stated a few decades ago and solved using classical (considered nowadays) methods, recently it has attracted more attention in the context of sparse modeling, where several notable contributions have been made. In the present manuscript, a new approach is considered in the framework of greedy algorithms. The noise is split into two components: a) the inlier bounded noise and b) the outliers, which are explicitly modeled by employing sparsity arguments. Based on this scheme, a novel efficient algorithm (Greedy Algorithm for Robust Denoising - GARD), is derived. GARD alternates between a least square optimization criterion and an Orthogonal Matching Pursuit (OMP) selection step that identifies the outliers. The case where only outliers are present has been studied separately, where bounds on the *Restricted Isometry Property* guarantee that the recovery of the signal via GARD is exact. Moreover, theoretical results concerning convergence as well as the derivation of error bounds in the case of additional bounded noise are discussed. Finally, we provide extensive simulations, which demonstrate the comparative advantages of the new technique.

1 Introduction

The notion of *robustness*, i.e., the efficiency of a method to solve a learning task from data, which have been contaminated by large values of noise, has occupied the scientific community for over half a century [1, 2]. Regardless the type of the problem, e.g., classification or regression, the goal is to identify observations that have been hit by large values of noise, known as *outliers*, and be removed from the training data set. Over the years, many authors have tried to state definitions to identify a data point as an outlier. A few typical characterizations follow:

- “An outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism” (Hawkins, 1980) [3].
- “An outlier is an observation which appears to be inconsistent with the remainder of the data set” (Barnet and Lewis, 1994) [4].
- “An outlier is an observation that lies outside the overall pattern of a distribution” (Moore and McCabe, 1999) [5].
- “An outlier in a set of data is an observation or a point that is considerably dissimilar or inconsistent with the remainder of the data” (Ramasmaw, 2000) [6].
- “Outliers are those data records that do not follow any pattern in an application” (Chen, 2002) [7, 8, 9].

In this paper, we focus on solutions of the linear regression problem in the presence of outliers. In such tasks, classic estimators, e.g., the Least-Squares, are known to fail [10]. This problem has been addressed since the 1950’s, in [1, 2] and actually solved more than two decades later, in [10, 11, 12], leading to the development of a new field in Statistics, known as *Robust Statistics*.

The variety of methods that have been developed to handle outliers can be classified into two major directions. The first one includes methods that rely on the use of *diagnostic tools*, where one tries first to delete the outliers and then to fit the “good” data by Least-Squares. The second direction, i.e., *robust*

*geopapag@di.uoa.gr

[†]panbouboulis@gmail.com

[‡]stheodor@di.uoa.gr

analysis, includes methods that firstly fit the data, using a rough approximation, and then exploit the original estimation to identify the outliers as those points which possess large residuals. Both approaches have a long history. Methods developed under the Robust statistics framework, consist of combinatorial optimization algorithms like Hampel's Least Median of Squares Regression (LMedS) [13] (p.16), Fischler's and Bolles's Random Sample Consensus (RANSAC) [14], as well as Rousseeuw's Least Trimmed Squares (LTS) [15, 13]. Combinatorial optimization methods seemed to perform well at that time, although they were never adopted by the community. Nowadays, where the size of the training data set can be very large such techniques are prohibited. In contrast, the desire for lower complexity efficient algorithms has constantly been rising. One of the pioneering research works at that time, was the development of Huber's M-est [16, 15, 10], a method that belongs to the category of robust analysis. M-est provides good estimates, without a heavy computational cost, using robust functions of the residual norm (instead of the square function), in order to penalize large values of the residual.

The development of methods in the spirit of robust analysis, owes a lot to the emergence of *sparse* optimization methods, during the past decade. Sparsity-aware learning and related optimization techniques have been at the forefront of the research in signal processing, encompassing a wide range of topics, such as compressed sensing, denoising and signal approximation techniques [17, 18, 19, 20, 21, 22]. There are two major paths, towards modeling sparse vectors/signals. The first one, focuses on minimizing the ℓ_0 (pseudo)-norm of a vector which equals the number of non-zero coordinates of a vector (this is a non-convex function), whereas the second one employs the ℓ_1 norm, the closest convex relaxation to the ℓ_0 (pseudo)-norm to regularize the Least-Squares cost function. Both methods have been shown to generate sparse solutions.

The family of algorithms that have been developed to address problems involving the ℓ_0 (pseudo)-norm, comprises *greedy* algorithms, which have been shown to provide the solution of the related minimization task, under certain reasonable assumptions, [23, 24, 25, 26, 27]. Even though, in general, this is an NP-hard problem, it has been shown that such methods can efficiently recover the solution in polynomial time. On the other hand, the family of algorithms developed around the methods that employ the ℓ_1 norm, embraces convex optimization, which provide a broader set of tools and stronger guarantees for convergence [25, 17, 28, 18, 29, 22].

In the present paper, the robust linear regression problem is approached via a celebrated greedy algorithm, the so called Orthogonal Matching Pursuit (OMP). The main idea is to split the noise into two separate components; one that corresponds to the inlier bounded noise and the other to the outliers. Since the outlier part is not present in all samples, sparse modeling arguments are mobilized to model the outlier noise component. This concept has been also employed in [30, 31, 32]. The novelty of our approach lies in the different modeling compared to already established works, by treating the task in terms of the ℓ_0 minimization via greedy algorithmic concepts. A new algorithm has been derived which exhibits notable performance gains, both in terms of computational resources as well as in terms of quality of the recovered results. Moreover, theoretical results concerning the power of the method to recover the outliers as well as performance error bound are derived.

The paper is organized as follows: In Section 2, a brief overview of the various methods that address the robust linear regression task is given. Section 3 presents in details the proposed algorithmic scheme (GARD). The main theoretical results concerning GARD, including convergence, recovery of the support, recovery error, e.t.c. are included in section 4. To validate the proposed method, section 5 includes several experiments, which compare GARD with other existing techniques. It is shown that GARD offers improved recovery at a reduced computational requirements. Finally, Section 6 contains some concluding remarks.

Notations: Throughout this work, capital letters are employed to denote sets, e.g., S , where S^c and $|S|$ denote the complement and the cardinality of the S respectively. The set of integer numbers between 1 and n , i.e., $\{1, 2, \dots, n\}$, will be denoted as $1-n$. Bold capital letters denote matrices, e.g., \mathbf{X} , while bold lowercase letters are reserved for vectors, e.g., $\boldsymbol{\theta}$. The symbol \cdot^T denotes the transpose of the respective matrix/vector. The i -th column of matrix \mathbf{X} is denoted by \mathbf{x}_i and the i -th element of vector $\boldsymbol{\theta}$ is denoted by θ_i . The matrix \mathbf{X}_S is the matrix \mathbf{X} restricted over the set S , i.e., the matrix which comprises of the columns of \mathbf{X} , whose indices belong to the ordered index set $S = \{j_1 < \dots < j_s\}$. Moreover, the identity matrix of dimension n will be denoted as \mathbf{I}_n , the zero matrix of dimension $n \times m$, as $\mathbf{O}_{n \times m}$, the vector of zero elements of appropriate dimension as $\mathbf{0}$ and columns of matrix \mathbf{I}_n restricted over the set S , as \mathbf{I}_S . If $\mathbf{v} \in \mathbf{R}^n$ is an s -sparse vector over the support set $S \subset 1-n$, with $|S| = s$, then we denote as $[\mathbf{v}]_S \in \mathbf{R}^s$, or \mathbf{v}_S for short, the vector which contains only the s non-zero entries of \mathbf{v} , i.e., $\mathbf{v}_S = \mathbf{I}_S^T \mathbf{v}$. For example, if $\mathbf{v} = (5, 0, 0, 2, 0)^T$, then $\mathbf{v}_{\{1,4\}} = (5, 2)^T$. Moreover, one can easily show that $\mathbf{I}_S \mathbf{v}_S = \mathbf{v}$. Finally, we use a linear tensor defined as $F_{S'}(\boldsymbol{\alpha}) = \mathbf{I}_{S'} \mathbf{I}_{S'}^T \boldsymbol{\alpha}$, for any vector $\boldsymbol{\alpha} \in \mathbf{R}^n$ over the index set $S' \subseteq 1-n$, which has identical coordinates with $\boldsymbol{\alpha}$ in all indices of S' and zero everywhere else. Finally, it is obvious that for the sparse vector \mathbf{v} with support set S , $F_S(\mathbf{v}) = \mathbf{I}_S \mathbf{I}_S^T \mathbf{v} = \mathbf{v}$.

2 Preliminaries and related work

In a typical linear regression task, we are interested in estimating the linear relation between two variables, $\mathbf{x} \in \mathbf{R}^m$ and $y \in \mathbf{R}$, i.e., $y = \mathbf{x}^T \boldsymbol{\theta}$, when several noisy instances, i.e., $\{(y_i, \mathbf{x}_i), i = 1, \dots, n\}$, are known. In this context, we usually adopt the following (regression) modeling

$$y_i = \mathbf{x}_i^T \boldsymbol{\theta}_0 + e_i, \quad i = 1, \dots, n, \quad (1)$$

where e_i is some observation noise. Hence, our goal is to estimate $\boldsymbol{\theta}_0 \in \mathbf{R}^m$ from the given training dataset of n observations. In matrix notation, Eq. (1) can be rewritten as follows:

$$\mathbf{y} = \mathbf{X} \boldsymbol{\theta}_0 + \mathbf{e}, \quad (2)$$

where $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$, $\mathbf{e} = (e_1, e_2, \dots, e_n)^T$ and $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^T \in \mathbf{R}^{n \times m}$.

As it is common in regression analysis, we consider that the number of observations exceeds the number of unknowns, i.e., $n > m$. In order to seek for a solution, we should assume that \mathbf{X} is a full rank matrix, i.e. $\text{rank}(\mathbf{X}) = m$. If the noise is i.i.d Gaussian, the most common estimator, which is statistically optimal (BLUE), is the Least-Squares (LS) one. However, this is not the case in the presence of outliers or when the noise distribution exhibits long tails. In the following, we will give a brief overview of the algorithmic schemes that has been proposed to deal with the aforementioned problem. These schemes can be classified into two major categories, those that apply a weighted scheme to penalize the largest residuals and those that apply sparse modeling.

It should also be noted, that for the case where $n < m$ (underdetermined system of linear equations¹), an additional condition/constraint should be imposed, if one wishes to recover the unknown vector. One of the most common features lately, is to impose sparsity constraints on $\boldsymbol{\theta}_0$. However, in such a case, the task breaks down to a classical sparse model for the combined matrix $[\mathbf{X} \quad \mathbf{I}]$, which has been extensively studied over the last years. Thus, any sparse related algorithm, e.g., ADMM for solving the LASSO formulation, OMP or any other greedy approach method, is rendered suitable for performing the estimation. Hence, the study of this case is considered as trivial.

2.1 Penalizing Large Residuals

Both methods of this group, attempt to estimate $\boldsymbol{\theta}_0$, based on equation (2).

- **M-estimators (M-est) [16]:**

In M-est a robust cost function ρ (satisfying certain properties) of the residual error $r_i = y_i - \mathbf{x}_i^T \boldsymbol{\theta}$, $i = 1, 2, \dots, n$, is minimized, so that $\boldsymbol{\theta}_* = \arg \min_{\boldsymbol{\theta}} \sum_{i=1}^n \rho(r_i)$. Differentiation with respect to $\boldsymbol{\theta}$ leads to $\sum_{i=1}^n \psi(r_i) \mathbf{x}_i^T = \mathbf{0}$, where $\psi = \rho'$. If we define $w(r) = \psi(r)/r$, $w_i = w(r_i)$, then the system of normal equations is cast as $\sum_{i=1}^n w_i r_i \mathbf{x}_i^T = \mathbf{0}$. This is the basic version of M-est, although several variations exist. In our experiments, we have used a scaling parameter $\hat{\sigma}$, computed at each step, and defined $\rho := \rho(r_i/\hat{\sigma})$. Consequently, another way to interpret M-est, is by solving a Weighted Least-Squares problem,

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^n w_i r_i^2 \Leftrightarrow \min_{\boldsymbol{\theta}} \|\mathbf{W}_{\mathbf{r}}^{1/2} (\mathbf{y} - \mathbf{X} \boldsymbol{\theta})\|_2^2. \quad (3)$$

To solve (3) the *Iteratively reweighted least squares* (IRLS) algorithmic scheme is employed, where the diagonal weight matrix $\mathbf{W}_{\mathbf{r}}$ assigns the weights, with values depending on the Robust function selected (Huber's, Tukey's biweight, Hampel's, Andrews). Notice that if $\mathbf{W}_{\mathbf{r}} = \mathbf{I}_n$, the scheme performs a classic Least-Squares. Many improved variants can be found in the literature. For more details, the interested reader is referred to [16, 13, 15].

- **Robust Orthogonal Matching Pursuit (ROMP) [33]:**

The method is based on the M-est and the popular OMP algorithm, which was studied in [24, 34, 23]. However, the key aspect of the algorithm, which is also the feature that introduces robustness, is the execution of a weighted Least Squares step (M-est), instead of an ordinary one, each time the support set is augmented with an atom. Although a variety of termination criteria exist, we let the algorithm terminate, as soon as the length of the residual vector drops below a predefined threshold. The method is summarized as follows: *Initialization*: $k := 0$, $\Omega^0 := \emptyset$, $\boldsymbol{\theta}^{(0)} = \mathbf{0}$, $\mathbf{r}^{(0)} = \mathbf{y}$.

Main iteration

¹An infinite number of solutions exist.

Step 1 - Initialization:

$$k := k + 1, \hat{\sigma} = MAD(\mathbf{r}^{(k-1)}), \mathbf{r}_{\psi}^{(k-1)} = \psi(\mathbf{r}^{(k-1)}/\hat{\sigma}).$$

Step 2 - Atom selection:

$$i_k := \arg \max \left| \mathbf{X}^T \mathbf{r}_{\psi}^{(k-1)} \right|, \Omega^k := \Omega^{k-1} \cup i_k.$$

Step 3 - Solution:

$$\boldsymbol{\theta}^{(k)} := \arg \min_{\boldsymbol{\theta}} \|\mathbf{W}_{\mathbf{r}}^{1/2}(\mathbf{y} - \mathbf{X}_{\Omega^k} \boldsymbol{\theta})\|_2^2, \quad (4)$$

$$\mathbf{r}^{(k)} = \mathbf{y} - \mathbf{X}_{\Omega^k} \boldsymbol{\theta}^{(k)}.$$

The main procedure begins with the computation of $\hat{\sigma} = MAD(\mathbf{r})^2$ and the residual pseudo-values \mathbf{r}_{ψ} , which are then used for selecting the atom of matrix \mathbf{X} , that is most correlated to the residual pseudo-values (Step 2). Finally, it should be noted, that ψ is a robust function (as in M-est), that also assigns the weights of matrix $\mathbf{W}_{\mathbf{r}}$, required for solving the weighted Least Squares step, in (4). Here, the difference to (3), is that at each k step, \mathbf{X}_{Ω^k} includes only the columns of \mathbf{X} that have been selected until the current step. Unfortunately, no theoretical justifications have been made, either on the selection of the atom based on the residual pseudo-values or on the iterative employment of the M-est.

2.2 Sparse outlier modeling

For all of the following methods, a different model is adopted. To this end, assume, that the outlier noise values are significantly fewer (i.e., sparse) compared to the size of the input data. Thus, a familiar technique, is to express the noise vector as a sum of two independent components, $\mathbf{e} = \mathbf{u} + \boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is assumed to be the dense inlier noise vector of energy ϵ_0 and $\mathbf{u} \in \mathbf{R}^n$ the sparse outlier noise vector with support set, T , and cardinality $|T| \leq s \ll n$. The support set is defined as the set of indices $i \in \{0, \dots, n\}$ that satisfy $u_i \neq 0$. Hence, equation (2) can be recast as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}_0 + \mathbf{u}_0 + \boldsymbol{\eta}. \quad (5)$$

As we would like to minimize the number of outliers in (5), the associated optimization problem becomes:

$$\min_{\boldsymbol{\theta}, \mathbf{u}} \|\mathbf{u}\|_0, \text{ s.t. } \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta} - \mathbf{u}\|_2 \leq \epsilon_0. \quad (6)$$

However, in general, the task in (6) is a combinatorial problem. Hence, many authors propose to relax the ℓ_0 with the ℓ_1 norm, using a similar formulation:

$$\min_{\boldsymbol{\theta}, \mathbf{u}} \|\mathbf{u}\|_1, \text{ s.t. } \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta} - \mathbf{u}\|_2 \leq \epsilon_0, \quad (7)$$

This has the advantage of transforming (6) to a convex problem, which can be solved using a variety of methods.

- **LASSO formulation for robust denoising [35, 36, 31]:**

The Alternating Direction Method of Multipliers (ADMM) is a technique for solving the Lagrangian form of (7), for appropriate multiplier values $\lambda > 0$ (Generalized Lasso form):

$$\mathbf{w}_* := \arg \min_{\mathbf{w}} \{(1/2)\|\mathbf{y} - \mathbf{A}\mathbf{w}\|_2^2 + \lambda\|\mathbf{F}\mathbf{w}\|_1\}, \quad (8)$$

where $\mathbf{w} = (\boldsymbol{\theta}, \mathbf{u})^T$, $\mathbf{A} = [\mathbf{X} \ \mathbf{I}_n]$, $\mathbf{F} = [\mathbf{O}_{n \times m} \ \mathbf{I}_n]$ (for $\mathbf{F} = \mathbf{I}_{n+m}$, we have the standard Lasso form). The ADMM method was studied in the 70's and 80's, as a good alternative to penalty methods, although it was established as a method to solve partial differential equations, [37, 38].

- **Second Order Cone Programming (SOCP):**

Problem (7) is also known as *Robust Regression Basis Pursuit*-(BPRR) and it can be reformulated as a Second Order Cone Programming (SOCP) task [39, 40]:

$$\begin{aligned} & \mathbf{w}_* := \arg \min_{\mathbf{w}} \mathbf{g}^T \mathbf{w}, \\ \text{s.t. } & \mathbf{H}^T \mathbf{w} \geq \mathbf{0}, \mathbf{y} - \mathbf{R}\mathbf{w} \in \mathcal{C}_{\epsilon_0}^{n+1} \end{aligned} \quad (9)$$

²Median Absolute Deviation $MAD(\mathbf{x}) = \text{median}_i(|x_i - \text{median}_i(x_i)|)$.

where $\mathbf{g} = (\mathbf{0}, \mathbf{0}, \mathbf{1})^T \in \mathbf{R}^{m+2n}$, $\mathbf{w} = (\boldsymbol{\theta}, \mathbf{u}, \mathbf{s})^T$,

$$\mathbf{H} = \begin{bmatrix} \mathbf{O}_{m \times n} & \mathbf{O}_{m \times n} \\ -\mathbf{I}_n & \mathbf{I}_n \\ \mathbf{I}_n & \mathbf{I}_n \end{bmatrix}, \quad \mathbf{R} = [\mathbf{X} \ \mathbf{I}_n \ \mathbf{O}_{n \times n}]$$

and $\mathcal{C}_{\epsilon_0}^{n+1}$ is the unit second order (convex) cone of dimension $n + 1$.

- **Sparse Bayesian Learning (SBL)** [41, 42, 30, 22]:

Another path that has been exploited in the respective literature is to use Sparse Bayesian Learning techniques [32, 30]. Assume that u_i is a random variable with prior distribution $u_i \sim \mathcal{N}(0, \gamma_i)$, where γ_i is the hyperparameter that controls the variance of each u_i that has to be learnt. If $\gamma_i = 0$, then $u_i = 0$, i.e., no outlier exists on this index. In contrast, a positive value of γ_i , results in an outlier in the measurement i . To estimate the regression coefficients, we jointly find

$$(\boldsymbol{\theta}_*, \boldsymbol{\gamma}_*, \sigma_*^2) = \arg \max_{\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma^2} P(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma^2), \quad (10)$$

where $\boldsymbol{\gamma} := (\gamma_1, \gamma_2, \dots, \gamma_n)^T$ and $\eta_i \sim \mathcal{N}(0, \sigma^2)$. The posterior estimation of \mathbf{u} , follows, from:

$$\mathbf{u}_* = E[\mathbf{u} | \mathbf{X}, \boldsymbol{\theta}_*, \boldsymbol{\gamma}_*, \sigma_*^2]. \quad (11)$$

3 Greedy Algorithm for Robust Denoising (GARD)

The goal of the proposed algorithmic scheme is to solve problem (6) using the split noise model described in (5) and it is designed along the celebrated Orthogonal Matching Pursuit rationale. It should be noted that ROMP, which also employs OMP's selection technique, is quite different from our approach. ROMP is mainly based on the M-est algorithm, while the proposed scheme, in contrast to other methods, tackles directly problem (6) and alternates between a least squares minimization task and an OMP selection technique. It can be easily seen that (6) can also be cast as:

$$\min_{\boldsymbol{\theta}, \mathbf{u}} \|\mathbf{u}\|_0, \quad \text{s.t.} \quad \left\| \mathbf{y} - \mathbf{A} \begin{pmatrix} \boldsymbol{\theta} \\ \mathbf{u} \end{pmatrix} \right\|_2 \leq \epsilon_0, \quad (12)$$

where $\mathbf{A} = [\mathbf{X} \ \mathbf{I}_n]$. Following OMP's rationale, at each iteration step, GARD estimates the solution, i.e., $\mathbf{z}_*^{(k)} = (\boldsymbol{\theta}_*^{(k)}, \mathbf{u}_*^{(k)})^T \in \mathbf{R}^{m+k}$ (for step $k = 0$ no outlier estimates exist and $\mathbf{z}_*^{(0)} \in \mathbf{R}^m$), using a Least-Squares criterion (i.e., $\min_{\boldsymbol{\theta}, \mathbf{u}} \|\mathbf{y} - \mathbf{A}(\boldsymbol{\theta}, \mathbf{u})^T\|^2$). In the following iterations GARD selects the observation which is the furthest away from the solution, using OMP's selection rationale (based on correlation). Hence, GARD restricts the selection over atoms of the second half of matrix \mathbf{A} , i.e., matrix $\mathbf{I}_n = [\mathbf{e}_1 \ \mathbf{e}_2 \ \dots \ \mathbf{e}_n]$, where \mathbf{e}_i are the vectors of the standard basis of \mathbf{R}^n .

To be more specific, at the first step the algorithm computes the initial Least-Squares solution disregarding the presence of outliers, i.e., $\boldsymbol{\theta}_* = \arg \min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2$, and the initial residual $\mathbf{r}^{(0)} = \mathbf{y} - \mathbf{X}\boldsymbol{\theta}_*$. Moreover, the set of the so called *active columns* is initialized to include all the columns of \mathbf{X} . Then, the main iteration cycle begins. At each step, GARD can be divided into two parts:

- Firstly, the greedy selection step is performed, i.e., the column vector from matrix \mathbf{I}_n that is more correlated with the latest residual is selected and the set of active columns of \mathbf{A} (i.e., the set of columns that have already been selected) is augmented by that column. The correlation is measured with respect to the angle, which in turn leads to the maximization of $|\langle \mathbf{r}^{(k)}, \mathbf{e}_i \rangle| = |r_i^{(k)}|$ for an index $i \in J = 1-n$.
- Next, a Least-Squares solution step is performed and the new residual is computed.

This procedure is repeated until the residual drops below a specific predefined threshold, as described in details in Algorithm 1. In order to avoid any confusion, we should also emphasize that even though both the sets J and S_{inac} correspond to the same orthonormal vectors \mathbf{e}_i of matrix \mathbf{I}_n , they should not be regarded as equal, since J includes indices from \mathbf{I}_n , whereas S_{inac} includes indices from the second half of the augmented matrix \mathbf{A} . Consequently, since $\mathbf{r}^{(k-1)} \in \mathbf{R}^n$ and the index selected, i.e., $j_k \in S_{inac}$, exceeds the dimensions of \mathbf{I}_n , the index $j - m$ is used for the elements of $\mathbf{r}^{(k-1)}$ in order to include indices that belong to the set J . For instance, if the largest component of $|\mathbf{r}^{(k-1)}|$ is the 5-th one, this leads to the fact that $j_k - m = 5 \in J$ and $j_k = m + 5 \in S_{inac}$. As it will be shown, the improved performance of the proposed scheme is due to

Algorithm 1 : Greedy Algorithm for Robust Denoising (GARD)

Input: $\mathbf{X}, \mathbf{y}, \epsilon_0$
Output: $\mathbf{z}_* = (\boldsymbol{\theta}_*, \mathbf{u}_*)^T$
 Initialization: $k := 0$
 $S_{ac} = \{1, 2, \dots, m\}, S_{inac} = \{m+1, \dots, m+n\}$
 $\mathbf{A}_{ac}^{(0)} = \mathbf{X}$
Solution*: $\mathbf{z}_*^{(0)} := \arg \min_{\mathbf{z}} \|\mathbf{y} - \mathbf{A}_{ac}^{(0)} \mathbf{z}\|_2^2$
 Initial Residual: $\mathbf{r}^{(0)} = \mathbf{y} - \mathbf{A}_{ac}^{(0)} \mathbf{z}_*^{(0)}$
while $\|\mathbf{r}^{(k)}\|_2 > \epsilon_0$ **do**
 $k := k + 1$
 Selection: $j_k := \arg \max_{j \in S_{inac}} |r_{j-m}^{(k-1)}|$
 Update Support: $S_{ac} := S_{ac} \cup \{j_k\}, S_{inac} = S_{ac}^c, \mathbf{A}_{ac}^{(k)} = [\mathbf{A}_{ac}^{(k-1)} \quad \mathbf{e}_{j_k}]$
 Update Solution:** $\mathbf{z}_*^{(k)} := \arg \min_{\mathbf{z}} \|\mathbf{y} - \mathbf{A}_{ac}^{(k)} \mathbf{z}\|_2^2$
 Update Residual: $\mathbf{r}^{(k)} = \mathbf{y} - \mathbf{A}_{ac}^{(k)} \mathbf{z}_*^{(k)}$
end while

the orthogonality between the columns of \mathbf{I}_n (standard Euclidean basis). The complexity of the algorithm is $O((m+k)^3/3 + n(m+k)^2)$ at each k step, making it unattractive when the dimension of the unknown vector is large. However, since at each step the method solves a standard Least-Squares problem, the complexity could be further reduced using *Cholesky* decomposition, *QR* factorization or the *matrix inversion Lemma*. For details on those implementations of the classic OMP, read [43]. Playing with all schemes, we found that in our case the most efficient implementation was the Cholesky decomposition, as described below:

- Replace the initial solution step $k := 0$, of Algorithm 1, with:

Solution*:
 Factorization step: $\mathbf{W}_0 = \mathbf{X}^T \mathbf{X}$ as $\mathbf{W}_0 = \mathbf{L}_0 \mathbf{L}_0^T$.
 Solve $\mathbf{L}_0 \mathbf{L}_0^T \mathbf{z} = \mathbf{X}^T \mathbf{y}$ using:

 - forward substitution $\mathbf{L}_0 \mathbf{q} = \mathbf{X}^T \mathbf{y}$
 - backward substitution $\mathbf{L}_0^T \mathbf{z}_*^{(0)} = \mathbf{q}$.
- Replace the update solution step $k := k+1$, of Algorithm 1, with:

Update Solution:**
 Compute \mathbf{v} such that: $\mathbf{L}_{k-1} \mathbf{v} = \mathbf{A}_{ac}^{(k-1)^T} \mathbf{e}_{j_k}$
 Compute: $b = \sqrt{1 - \|\mathbf{v}\|_2^2}$
 Matrix Update: $\mathbf{L}_k = \begin{pmatrix} \mathbf{L}_{k-1} & \mathbf{0} \\ \mathbf{v}^T & b \end{pmatrix}$
 Solve $\mathbf{L}_k \mathbf{L}_k^T \mathbf{z} = \mathbf{A}_{ac}^{(k)^T} \mathbf{y}$ using:

 - forward substitution $\mathbf{L}_k \mathbf{p} = \mathbf{A}_{ac}^{(k)^T} \mathbf{y}$
 - backward substitution $\mathbf{L}_k^T \mathbf{z}_*^{(k)} = \mathbf{p}$.

This modification leads to a reduction of the cost by an order of magnitude, for the main iteration steps. Analytically, the cost at the initial factorization, plus the cost for the forward and backward substitution, is $O(m^3/3 + nm^2)$. At each next step, neither inversion nor factorization is required. The lower triangular matrix \mathbf{L}_k is updated, only with a minimal cost of square-dependence. Furthermore, the cost required for solving the linear system using forward and backward substitution at each next k step is $O((m+k)^2 + 2n(m+k))$. Thus, the *total complexity* of the efficient (via the Cholesky decomposition) GARD implementation is $O(m^3/3 + k^3/2 + (n+3k)m^2 + 3kmn)$.

Remark 1. The algorithm begins with a Least-Squares solution to obtain $\mathbf{z}_*^{(0)}$. Thus, if no outliers exist, GARD solves the standard Least-Squares problem; it provides the maximum likelihood estimator (MLE), when the noise is Gaussian.

Remark 2. Since GARD solves a Least-Squares problem at each step, the new residual, $\mathbf{r}^{(k)}$, is orthogonal to each column that participates in the representation, i.e., $\langle \mathbf{r}^{(k)}, \mathbf{e}_{j_k} \rangle = r_{j_k}^{(k)} = 0, \forall k = 1, 2, \dots$. Thus, column \mathbf{e}_{j_k} of matrix \mathbf{I}_n , cannot be reselected.

Remark 3. Considering the complexity of the efficient implementation of GARD, the algorithm speeds up in cases where the fraction of the outliers is very low, i.e., the outlier vector is very sparse ($k \ll n$).

Remark 4. Matrix $\mathbf{A}_{ac}^{(k)} = [\mathbf{A}_{ac}^{(k-1)} \mathbf{e}_{j_k}]$ could also be cast as $\mathbf{A}_{ac}^{(k)} = [\mathbf{X} \mathbf{I}_{S_k}]$, where $S_k = \{j_1, j_2, \dots, j_k\}$, is the set of columns selected at the current step, i.e., the support set of our sparse estimate. Thus, the estimated support should not be confused with the set of active columns S_{ac} that participate in the representation of \mathbf{y} .

Remark 5. The proposed scheme should not be confused with other OMP-based schemes, such as Robust OMP in [33]; although both are OMP-based, they perform in a distinctive manner and for dissimilar purposes. As both the selection step, as well as the minimization step work quite different, GARD selects a column, based on the residual and performs a classic Least Squares procedure, whereas ROMP selects a column based on the residual pseudo-values and then solves a weighted Least Squares minimization problem.

4 Theoretical results

This section is devoted to study the main properties of GARD. Firstly, the convergence properties of the proposed scheme are derived. In the sequel, it is shown that GARD can recover the exact solution, under certain assumptions, in the presence of outlier noise only. Finally, for the case of both inlier and outlier noise, bounds for the recovery of the sparse outlier support, as well as the reconstruction error are presented.

4.1 General results

Lemma 1. At every $k \leq n - m$ step, GARD selects a column vector \mathbf{e}_{j_k} from matrix \mathbf{I}_n , that is linearly independent of all the column vectors in matrix $\mathbf{A}_{ac}^{(k-1)}$. Hence, $\mathbf{A}_{ac}^{(k)}$ has full rank and the solution to the Least-Squares problem at each step is unique.

Proof. The proof relies on mathematical induction. At the initial step, matrix $\mathbf{A}_{ac}^{(0)} = \mathbf{X}$ has been assumed to be full rank, hence the solution of the Least-Squares problem is unique. Suppose that at $k - 1$ step ($k \in \mathbf{N}^*$), matrix $\mathbf{A}_{ac}^{(k-1)}$ is full rank, hence let $\mathbf{z}_*^{(k-1)}$ denote the unique solution of the Least-Squares problem and $\mathbf{r}^{(k-1)} = \mathbf{y} - \mathbf{A}_{ac}^{(k-1)} \mathbf{z}_*^{(k-1)}$ the residual at the current step. Assume that at the k -th step, the j_k -th column of matrix \mathbf{I}_n is selected from the set S_{inac} . We will prove that the columns of the augmented matrix at this step, i.e., the columns of matrix $\mathbf{A}_{ac}^{(k)} = [\mathbf{A}_{ac}^{(k-1)} \mathbf{e}_{j_k}]$, are linearly independent. Since $j_k := \arg \max_{j \in S_{inac}} |r_j^{(k-1)}|$, we have that $r_{j_k}^{(k-1)} \neq 0$ (otherwise either this wouldn't have been selected or the residual vector would be equal to zero). Suppose that, the columns of matrix $\mathbf{A}_{ac}^{(k)}$ are linearly dependent, i.e., there exists $\mathbf{a} \neq \mathbf{0}$, such that $\mathbf{e}_{j_k} = \mathbf{A}_{ac}^{(k-1)} \mathbf{a}$ and let $\tilde{\mathbf{z}}^{(k-1)} = \mathbf{z}_*^{(k-1)} + r_{j_k}^{(k-1)} \mathbf{a}$. Thus, we have

$$\begin{aligned} \|\tilde{\mathbf{r}}^{(k-1)}\|_2 &= \|\mathbf{y} - \mathbf{A}_{ac}^{(k-1)} \tilde{\mathbf{z}}^{(k-1)}\|_2 = \\ &= \|\mathbf{y} - \mathbf{A}_{ac}^{(k-1)} \mathbf{z}_*^{(k-1)} - r_{j_k}^{(k-1)} \mathbf{A}_{ac}^{(k-1)} \mathbf{a}\|_2 = \\ &= \|\mathbf{r}^{(k-1)} - r_{j_k}^{(k-1)} \mathbf{e}_{j_k}\|_2 < \|\mathbf{r}^{(k-1)}\|_2, \end{aligned}$$

which contradicts the fact that the residual of the Least-Squares solution attains the smallest norm. Thus, all the selected columns of matrix $\mathbf{A}_{ac}^{(k)}$ are linearly independent. \square

Theorem 1. The norm of the residual vector

$$\mathbf{r}^{(k)} = \mathbf{y} - \mathbf{A}_{ac}^{(k)} \mathbf{z}_*^{(k)}$$

in GARD is strictly decreasing. Moreover, the algorithm will always converge.

Proof. Let $\mathbf{z}_*^{(k-1)}$ be the unique Least-Squares solution (Lemma 1) and $\mathbf{r}^{(k-1)} = \mathbf{y} - \mathbf{A}_{ac}^{(k-1)} \mathbf{z}_*^{(k-1)}$ the respective residual at $k - 1$ step. At the next step, the algorithm selects the column j_k and augments matrix $\mathbf{A}_{ac}^{(k-1)}$, by column \mathbf{e}_{j_k} to form matrix $\mathbf{A}_{ac}^{(k)}$. Let $\mathbf{z}_*^{(k)}$ denote the unique solution of the Least-Squares problem at the k -th step (Lemma 1) and $\mathbf{r}^{(k)} = \mathbf{y} - \mathbf{A}_{ac}^{(k)} \mathbf{z}_*^{(k)}$ the respective residual. Consequently, one could define a cost function for every $\mathbf{z} \in \mathbf{R}^{m+k}$ at k step, as $P^{(k)}(\mathbf{z}) = \|\mathbf{y} - \mathbf{A}_{ac}^{(k)} \mathbf{z}\|_2$. Thus, we have that

$$\|\mathbf{r}^{(k)}\|_2 = P^{(k)}(\mathbf{z}_*^{(k)}) \leq P^{(k)}(\mathbf{z}), \quad (13)$$

for every $\mathbf{z} \in \mathbf{R}^{m+k}$. Now let $\mathbf{z}^{(k)} = (\mathbf{z}_*^{(k-1)}, r_{j_k}^{(k-1)})^T$, where $r_{j_k}^{(k-1)}$ is the j_k coordinate of the residual $\mathbf{r}^{(k-1)}$. Thus, we have that

$$\begin{aligned} P^{(k)}(\mathbf{z}^{(k)}) &= \|\mathbf{y} - \mathbf{A}_{ac}^{(k)} \mathbf{z}^{(k)}\|_2 \\ &= \|\mathbf{y} - \mathbf{A}_{ac}^{(k-1)} \mathbf{z}_*^{(k-1)} - r_{j_k}^{(k-1)} \mathbf{e}_{j_k}\|_2 \\ &= \|\mathbf{r}^{(k-1)} - r_{j_k}^{(k-1)} \mathbf{e}_{j_k}\|_2 < \|\mathbf{r}^{(k-1)}\|_2. \end{aligned} \quad (14)$$

Combining (13) and (14), we have that

$$\|\mathbf{r}^{(k)}\|_2 < \|\mathbf{r}^{(k-1)}\|_2. \quad (15)$$

Since $\mathbf{y} \in \mathbf{R}^n$, the residual equals zero, as soon as $n - m$ columns have been selected. However, since the noise bound is a positive value assumed to be known, the algorithm terminates at the first step $k < n - m$, where the residual's norm drops below ϵ_0 . \square

4.2 The presence of outliers only

The scenario where the signal is corrupted only by outliers is treated separately. In this case, we aim to solve the following ℓ_0 minimization problem:

$$\begin{aligned} \min_{\boldsymbol{\theta}, \mathbf{u}} \quad & \|\mathbf{u}\|_0 \\ \text{s. t.} \quad & \mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{u}, \end{aligned} \quad (16)$$

where \mathbf{X} is assumed to be a full column rank matrix (note that if \mathbf{X} has linearly dependent columns, there is not a unique solution for this problem). The general solution of (16) is an NP-hard task. However, under specific conditions, the problem can be solved efficiently using GARD, as it will be proved subsequently.

To simplify notation and reduce the size of the subsequent proofs, we orthonormalize \mathbf{X} by the reduced QR decomposition, i.e., $\mathbf{X} = \mathbf{Q}\mathbf{R}$, where \mathbf{Q} is a $n \times m$ matrix, whose columns form an orthonormal basis of the column space of \mathbf{X} (i.e., $\text{span}(\mathbf{X})$) and \mathbf{R} is a $m \times m$ upper triangular matrix. Since \mathbf{X} has full column rank, the decomposition is unique; moreover, matrix \mathbf{R} is invertible. Using this decomposition, the split noise modeling described in equation (5) can be written as

$$\mathbf{y} = \mathbf{Q}\mathbf{w}_0 + \mathbf{u}_0 + \boldsymbol{\eta}, \quad (17)$$

where $\mathbf{w}_0 = \mathbf{R}\boldsymbol{\theta}_0$. If \mathbf{w}_0 is recovered, the unknown vector $\boldsymbol{\theta}_0$ can also be recovered from $\boldsymbol{\theta}_0 = \mathbf{R}^{-1}\mathbf{w}_0$. Equation (17) plays a central role, as it describes the model that is adopted throughout this paper. In this section, however, we assume that only outlier noise exist, hence $\boldsymbol{\eta}$ is set to zero.

We are now in the position to express equation (17) (for $\boldsymbol{\eta} = \mathbf{0}$) as $\mathbf{y} = [\mathbf{Q} \ \mathbf{I}_n] \mathbf{z}'_0$, where $\mathbf{z}'_0 = (\mathbf{w}_0^T, \mathbf{u}_0^T)^T$. Since the vector \mathbf{u}_0 is s -sparse at most, the measurement vector could also be written as $\mathbf{y} = [\mathbf{Q} \ \mathbf{I}_S] \mathbf{z}_0$, where \mathbf{I}_S is the matrix containing column vectors from \mathbf{I}_n indexed by³ S and $\mathbf{z}_0 = (\mathbf{w}_0^T, [\mathbf{u}_0]_S^T)^T$, with $[\mathbf{u}_0]_S \in \mathbf{R}^s$ representing the reduced vector⁴, that contains only the non-zero entries of \mathbf{u}_0 .

We assume that the outlier vector is sparse over the support subset $S \subset 1:n$, with $|S| = s \ll n$ (i.e., $u_i = 0$, for all $i \notin S$ and $u_i \neq 0$ for $i \in S$) and that $s < n/2$ (in the case where $s \geq n/2$, the solution cannot be recovered [28]). Applying the QR decomposition, problem (16) could also be written as:

$$\begin{aligned} \min_{\mathbf{w}, \mathbf{u}} \quad & \|\mathbf{u}\|_0 \\ \text{s. t.} \quad & \mathbf{y} = \mathbf{Q}\mathbf{w} + \mathbf{u}, \end{aligned} \quad (18)$$

In the following, the notion of the *smallest principal angle* between subspaces is employed. Given the information concerning the index subset S (i.e., we assume that we know the support of the outliers), \mathbf{w} can be recovered, if and only if $[\mathbf{Q} \ \mathbf{I}_S]$ has full rank. The latter assumption can also be expressed in terms of the *smallest principal angle*, ω_S , between the subspace spanned by the columns of the regressor matrix, i.e., $\text{span}(\mathbf{Q})$ and the subspace spanned by the columns of \mathbf{I}_S , i.e., $\text{span}(\mathbf{I}_S)$.

Definition 1. Let δ_S be the smallest number that satisfies the inequality $|\langle \mathbf{w}, \mathbf{u} \rangle| \leq \delta_S \|\mathbf{w}\|_2 \|\mathbf{u}\|_2$, for all $\mathbf{w} \in \text{span}(\mathbf{Q})$ and $\mathbf{u} \in \text{span}(\mathbf{I}_S)$. Then $\omega_S = \arccos(\delta_S)$ is the smallest principle angle between the spaces $\text{span}(\mathbf{Q})$ and $\text{span}(\mathbf{I}_S)$.

³Recall that $\mathbf{I}_S = [\mathbf{e}_{j_1}, \dots, \mathbf{e}_{j_{|S|}}]$, where $i_1 < \dots < i_{|S|}$ are the indices of S .

⁴Recall that $\mathbf{u} = \mathbf{I}_S \mathbf{u}_S$.

Generalizing Definition 1, we can take

$$\delta_s = \max \left\{ \delta_S, \text{ for all } S \in \binom{1-n}{k}, k \leq s \right\}, \quad (19)$$

where $1-n = \{1, \dots, n\}$ and $\binom{1-n}{k}$ denotes the set of all possible k -combinations of $1-n$. Hence, we define the smallest principal angle between the regression subspace $\text{span}(\mathbf{Q})$ and all the at most s -dimensional outlier subspaces; i.e., the spaces $\text{span}(\mathbf{I}_S)$ for all possible combinations of S , such that $|S| \leq s$, as follows:

$$\omega_s = \arccos(\delta_s). \quad (20)$$

It can readily be seen that δ_s can be defined by employing only the value $k = s$ (instead of all $k \leq s$) and that for any $\mathbf{w} \in \text{span}(\mathbf{Q})$ and any at most s -sparse vector \mathbf{u} we have that

$$|\langle \mathbf{w}, \mathbf{u} \rangle| \leq \delta_s \|\mathbf{w}\|_2 \|\mathbf{u}\|_2. \quad (21)$$

Remark 6. The quantity $\delta_s \in [0, 1]$ (or equivalently $\omega_s \in [0^\circ, 90^\circ]$) is a measure of how well separated the regressor subspace is from all the s -dimensional outlier subspaces.

The following condition, also well known as the Restricted Isometry Property (R.I.P.) and plays a central role in sparse optimization methods.

Definition 2. For orthonormal matrix \mathbf{Q} , we define a constant μ_s , $s = 1, 2, \dots, N$, as the smallest number such that

$$(1 - \mu_s) \|\boldsymbol{\alpha}\|_2^2 \leq \|[\mathbf{Q} \mathbf{I}_S] \boldsymbol{\alpha}\|_2^2 \leq (1 + \mu_s) \|\boldsymbol{\alpha}\|_2^2. \quad (22)$$

In [44] (Lemma III.1), it has been proved that for orthonormal regressor matrix \mathbf{Q} , the smallest principal angle coincides with the R.I.P. constant defined, i.e., $\delta_s = \mu_s$, $s = 1, 2, \dots, n$. Finally, the following theorem ([28, 44]), guarantees uniqueness of the decomposition.

Theorem 2. Assume that the vector $\mathbf{y} \in \mathbf{R}^n$ can be decomposed as follows:

$$\mathbf{y} = \mathbf{Q} \mathbf{w}_0 + \mathbf{u}_0, \quad (23)$$

where $\mathbf{w}_0 \in \mathbf{R}^m$ and \mathbf{u}_0 is an at most s -sparse vector. If $\delta_{2s} < 1$ then this decomposition is unique.

One of the main theoretical results, established in this work is the following theorem, which guarantees the recovery of the support of the sparse vector, which in turn leads to the recovery of the exact solution for the case only outliers exist.

Theorem 3. Let \mathbf{X} be a full column rank matrix and assume that the measurement vector, $\mathbf{y} = \mathbf{X} \boldsymbol{\theta}_0 + \mathbf{u}_0$, has a unique decomposition, such that $\|\mathbf{u}_0\|_0 \leq s$ (at most s outliers exist in the \mathbf{y} variable). If

$$\delta_s < \sqrt{\frac{\min\{|u_i|, u_i \neq 0\}}{2\|\mathbf{u}_0\|_2}}, \quad (24)$$

where u_i are the elements of \mathbf{u}_0 , then GARD guarantees that the unknown vector $\boldsymbol{\theta}_0$ and the sparse outlier vector \mathbf{u}_0 are recovered without any error.

The proof of the theorem is found in Appendix A.

Remark 7. The condition under which the measurement vector \mathbf{y} can be uniquely decomposed into parts $\mathbf{y}_0 = \mathbf{X} \boldsymbol{\theta}_0 = \mathbf{Q} \mathbf{w}_0$ plus \mathbf{u}_0 , is given in Theorem 2 (see also [28, 44]).

Remark 8. The bound found in (24), has also a nice geometrical interpretation. The ratio $\min\{|u_i|, u_i \neq 0\} / \|\mathbf{u}_0\|_2$, corresponds to the cosine of the largest direction angle of vector \mathbf{u}_0 . Moreover, it can readily be seen that this ratio is no greater than 1, which means that the right hand side of (24) is bounded by $\sqrt{2}/2$. In other words, the condition of Theorem 3 forces ω_s to lie in the interval $(45^\circ, 90^\circ]$.

4.3 The presence of both inlier and outlier noise

The following results show that when GARD recovers the support of the outlier vector, the approximation error is relatively small.

Theorem 4. *Let \mathbf{X} be a full column rank matrix and assume that $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}_0 + \mathbf{u}_0 + \boldsymbol{\eta}$, such that $\|\mathbf{u}_0\|_0 \leq s$ (at most s outliers exist in the \mathbf{y} variable) and also $\|\boldsymbol{\eta}\|_2 \leq \epsilon_0$. If*

$$\delta_s < \sqrt{\frac{\min\{|u_i|, u_i \neq 0\} - (2 + \sqrt{6})\epsilon_0}{2\|\mathbf{u}_0\|_2}}, \quad (25)$$

where u_i are the elements of \mathbf{u}_0 , $d = \lceil \frac{n}{s} \rceil$, then GARD guarantees that the support of the sparse outlier vector \mathbf{u}_0 is recovered⁵.

The proof of the theorem is found in Appendix B.

Lemma 2. *Assume that there exist $0 \leq \delta_s < 1$, such that the R.I.P. condition holds. It stems directly that the smallest singular value σ_{\min} of the matrix $\Phi_{S_{ac}} = [\mathbf{Q} \ \mathbf{I}_S]$ is lower bounded by*

$$\sigma_{\min} \geq \sqrt{1 - \delta_s}. \quad (26)$$

Proof. Let \mathbf{v}_m be the eigenvector which is associated with the smallest singular value of $\Phi_{S_{ac}}$, then

$$\|\Phi_{S_{ac}} \mathbf{v}_m\|_2^2 = \sigma_{\min}^2 \|\mathbf{v}_m\|_2^2.$$

Since (22) holds for every vector, (26) follows. \square

Theorem 5. *In the case where GARD recovers the exact support of the sparse outlier vector \mathbf{u}_0 , it approximates the ideal solution $\boldsymbol{\theta}_0$, with estimate $\boldsymbol{\theta}_*$, acquiring an error*

$$\|\boldsymbol{\theta}_* - \boldsymbol{\theta}_0\|_2 \leq \frac{\epsilon_0}{\tau \sqrt{1 - \delta_s}}. \quad (27)$$

where τ is the smallest singular value of matrix \mathbf{X} .

Proof. The proof follows the same concepts as the stability result of Theorem 5.1 in [45]. Since the support set of the sparse estimate \mathbf{u}_* is also S , matrix $\Phi = [\mathbf{Q} \ \mathbf{I}_n]$ could also be written as $\Phi = [\Phi_{S_{ac}} \ \mathbf{I}_{S^c}]$. Thus, we have that

$$\mathbf{z}_* = \begin{pmatrix} \mathbf{w}_* \\ [\mathbf{u}_*]_S \end{pmatrix} := \arg \min_{\mathbf{z}} \|\mathbf{y} - \Phi_{S_{ac}} \mathbf{z}\|_2^2 = \Phi_{S_{ac}}^\dagger \mathbf{y},$$

where matrix $\Phi_{S_{ac}}^\dagger$ denotes the Moore-Penrose pseudoinverse of matrix $\Phi_{S_{ac}}$. Equation (17) (which is equivalent to (5)) could also be written in a more compact form as

$$\mathbf{y} = \Phi_{S_{ac}} \mathbf{z}_0 + \boldsymbol{\eta},$$

where $\mathbf{z}_0 = \begin{pmatrix} \mathbf{w}_0 \\ [\mathbf{u}_0]_S \end{pmatrix}$. Hence, we take

$$\mathbf{z}_* = \Phi_{S_{ac}}^\dagger \mathbf{y} = \mathbf{z}_0 + \Phi_{S_{ac}}^\dagger \boldsymbol{\eta}.$$

Finally,

$$\begin{aligned} \|\mathbf{z}_* - \mathbf{z}_0\|_2 &\leq \|\Phi_{S_{ac}}^\dagger \boldsymbol{\eta}\|_2 \leq \|\Phi_{S_{ac}}^\dagger\|_2 \cdot \|\boldsymbol{\eta}\|_2 \\ &\leq \sigma_{\min}^{-1} \epsilon_0 \leq \epsilon_0 / \sqrt{1 - \delta_s}, \end{aligned} \quad (28)$$

where we have also used that $\|\Phi_{S_{ac}}^\dagger\|_2$ is bounded, using the smaller singular value σ_{\min} of matrix $\Phi_{S_{ac}}$, as well as (26). The result follows from the fact that

$$\|\boldsymbol{\theta}_* - \boldsymbol{\theta}_0\|_2 \leq \|\mathbf{R}^{-1}\|_2 \|\mathbf{z}_* - \mathbf{z}_0\|_2,$$

where $\|\mathbf{R}^{-1}\|_2$ is the spectral norm of \mathbf{R}^{-1} equal to $\sigma_{\min}(\mathbf{R})^{-1}$. Since $\mathbf{X} = \mathbf{Q}\mathbf{R}$, the smallest singular value of \mathbf{R} equals⁶ the smallest singular value $\tau = \sigma_{\min}(\mathbf{X})$ of \mathbf{X} , thus the proof is complete. \square

⁵Recall the definition of ceiling, i.e., $\lceil x \rceil = \min\{n \in \mathbb{Z} \mid n \geq x\}$.

⁶Matrices \mathbf{X} and \mathbf{R} share the same singular values.

Algorithm	Sol. to problem	Complexity
(GARD)	(6) or (12)	$O(m^3/3 + k^3/2 + (n + 3k)m^2 + 3kmn)$, $k \ll n$
(M-est)	(3)	$O(m^3/3 + nm^2)/\text{step}$
(SOCP)	(9)	$O((n + m)^{2.5}n)$
(ADMM)	(8)	$O((n + m)^3/3 + n(n + m)^2)/\text{step}$
(SBL)	(10)	$O(m^3/3 + nm^2)/\text{step}$
(ROMP)	(4)	-

Table 1: For (GARD), k is the number of times the algorithm identifies an outlier. For (GARD) and (SOCP) total complexity is given. For the rest total complexity depends on the number of iterations for convergence.

Remark 9. *Let*

$$c = \sqrt{\frac{\min\{|u_i|, u_i \neq 0\} - (2 + \sqrt{6})\epsilon_0}{2\|\mathbf{u}_0\|_2}}.$$

Although, c is readily computed, recall that δ_s is not, since this constant encloses the combinatorial nature of the problem for all the possible subsets of cardinality at most s . As a consequence, inequalities (24), (25), (26) and (27), could not be verified in practice; nonetheless, they all serve significant theoretical purposes.

Remark 10. *Combining (27) with (25), we also have the following bound for the approximation of θ_0 :*

$$\|\theta_* - \theta_0\|_2 \leq \frac{\epsilon_0}{\tau\sqrt{1-c}}, \quad (29)$$

which due to its immediacy will be tested and verified later, in section 5. However, it is looser than that of (27).

Remark 11. *The bound c in (25), clearly depends on the sparsity level and values of the outlier vector, but also on the inlier noise bound ϵ_0 . Also notice, that $\epsilon_0 = 0$ leads to the bound of δ_s in the noiseless case, i.e., (24). Since in (25) two terms affect the bound, we cannot expect to recover the support exactly, in case both outlier noise of high density and heavy inlier noise exist. Such a scenario, would imply the bound on δ_s to be extremely tight and it is likely that it could not be satisfied. Finally, notice that $\min\{|u_i|, u_i \neq 0\}$ should be greater than $(2 + \sqrt{6})\epsilon_0$, if we would like (25) to be valid. This could also be considered as a measure, of when an error value is considered as an outlier.*

5 Experiments

The setup for each one of the pre-existing methods (see section 2), which are compared with GARD, are listed below:

- (M-est): In the following experiments, Tukey’s biweight (or bisquare) robust (but nonconvex) function has been employed. This is included in the MATLAB function “robustfit”. For σ , we have used the default parameter value setting (see [13, 15]).
- (SOCP): In order to solve the (SOCP) problem, we employed the MATLAB function “SeDuMi”, which is included in the optimization package “CVX” of Stanford University, (CVX RESEARCH: <http://cvxr.com/> (31/01/2014)). The input parameter for SeDuMi, is the bound of the inlier noise, used for the definition of the second order cone.
- (ADMM): For this method, parameter λ should be predefined. Furthermore, the parameter, ρ , that is used for the soft thresholding operator is also predefined (low) to $\rho_0 = 10^{-4}$ and adapts at each step via $\rho_i = \min\{5, 1.1\rho_{i-1}\}$. We have also employed a termination criterion, when the norm of the estimate undergoes changes from one step to the next, less than the predefined threshold of 10^{-4} .
- (SBL): Input parameters, $\sigma_{(0)}^2$, $\theta_{(0)}$ and $\gamma_{i_{(0)}}$ are initialized. Following [41, 42], we have also pruned the hyperparameters $\gamma_{i_{(k)}}$ from future iterations, if they become smaller than a predefined threshold (set low to 10^{-5}). Although the computational cost for Robust (SBL) is $O(m^3/3 + nm^2)$ per step, the total cost depends on other variables too; such are the number of hyperparameters, that are pruned from future iterations, as well as the number of iterations needed for convergence. This is also the case for other methods, too.

- (ROMP): The algorithm makes use of OMP’s main iteration loop; in the first iteration, just a single column from matrix \mathbf{X} participates in the M-est solution and each time the number of columns is augmented by one. Since the method solves an (IRLS) (or M-est at each step), instead of solving a Least-Squares problem restricted on the active set of columns, the complexity of the algorithm is not given in closed form. Once again, we have used Tukey’s biweight function (*robustfit*) with the default parameter settings, as in the M-est. The algorithm is terminated once the residual error drops below the bound of the inlier noise ϵ_0 .

In the experiments section, we have tested and analyzed the performance of each related algorithm. The experimental set up parallels that of [44]. Our data (y_i, \mathbf{x}_i) , $i = 1, 2, \dots, n$, $\mathbf{x}_i \in \mathbf{R}^m$ are generated via equation (5); for the case where no inlier noise exists, we have set $\boldsymbol{\eta} = \mathbf{0}$. The rows of matrix \mathbf{X} , i.e., \mathbf{x}_i ’s, are obtained by uniformly sampling an m -dimensional hypercube centered around the origin and $\boldsymbol{\theta}_0 \in \mathbf{R}^m$ are random vectors, with values chosen from the normal distribution with mean value 0 and standard deviation set to 5.

5.1 Mean-Square Error estimation

In the first experiment, we compared all methods with respect to the mean-square error (MSE), which is computed as the average, over 100 realizations at each outlier vector density, of the squared norm of the difference between the estimation vector $\boldsymbol{\theta}_*$ and the unknown vector $\boldsymbol{\theta}_0$. In parallel, we have also computed the average time (MT) (in sec) that each method requires to complete the estimation, for each outlier density. Aiming for more details, we have plotted the results in a logarithmic scale for each dimension of the unknown vector/signal $\boldsymbol{\theta}_0$ ($m = 50, 100, 170$).

Outlier values are equal to ± 25 , in s indices, uniformly sampled over n coordinates ($s < n$). Although, an outlier vector is sparse by definition, in some experiments we extended the density level, in order to test each method to its limits. The inlier noise vector has elements drawn from the standard Gaussian distribution, with $\sigma = 1$ and inlier noise bound ϵ_0 , which is assumed to be known.

The input parameter for GARD, SOCP and ROMP is the inlier noise bound ϵ_0 . For ADMM, the regularization parameter is set to $\lambda = 1.2$. Note that all methods were carefully tuned so that to optimize their performance. A major drawback of the SBL is its sensitivity to the choice of the initial values. Recall that this is a non-convex method, which cannot guarantee that the global minimum is attained for each dimension m , while the time needed for each implementation cannot be assured, since the number of iterations until convergence strongly depend on those parameters. Hence, for this method, random initialization was performed a number of times and the best solution was selected. Finally, the (M-est) does not require any predefined parameters.

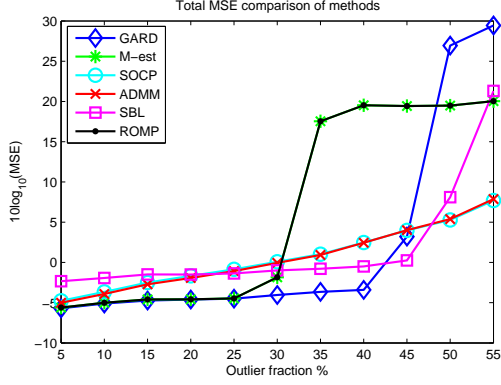
Figure ?? shows the MSE (in dBs), versus the fraction of the sparse outlier vector for various dimensions of the unknown vector $\boldsymbol{\theta}_0$. The MT, that is required for each algorithm to converge is shown in Figure ?? in logarithmic scale. Although complexity of each method is already addressed (table 1), in certain algorithms, the number of iterations until convergence greatly influences the required total implementation time. Observe that GARD attains the lowest MSE among the competitive methods for outlier fraction lower than 40%, 35% and 25% for dimension of the unknown vector $m = 50, 100, 170$, respectively. The performance of M-est and ROMP is also notable, since both methods also attain a low MSE. However, this is only possible for outlier fraction of less than 25%, 20% and 15% (MSE equal to that of GARD). In particular, we found that M-est and ROMP have identical performance, despite the fact that ROMP combines two methods, resulting to a higher computational cost.

It should also be noted, that in Figure ?? (b) and (c), all algorithms break their performance at lower outlier fractions with respect to GARD. However, the interesting *zone* of outlier vector density, in real time applications, is between 0% – 20% of the sample set, since greater percentages do not imply outlying values. Hence, GARD attains the lowest MSE within this sensitive zone. Finally, the experiments show that ADMM and SOCP attain a similar performance, as expected, due to the fact that they both address the same problem.

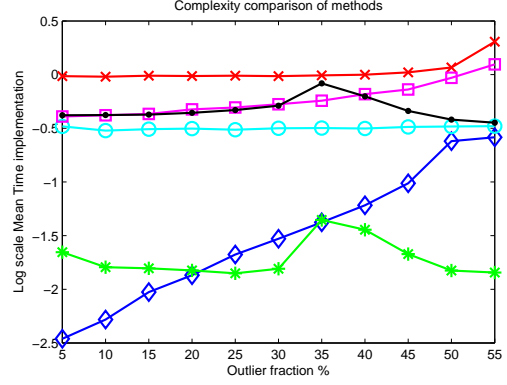
Besides its superior performance with respect to recovery error, GARD’s computational requirements remain low. As shown in Figure ??, GARD appears to have the lower computational cost among its competitors, for outlier fraction less than 20%.

5.2 Complexity evaluation for large data sets

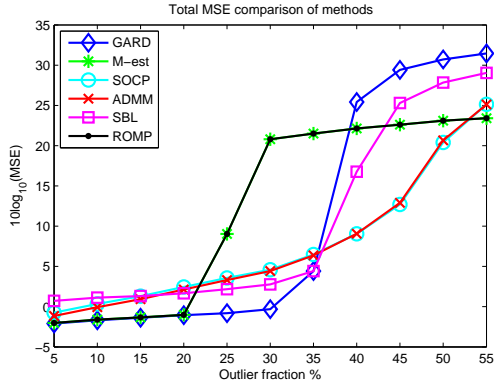
In the current section, we have attempted to evaluate the performance of the most computationally efficient methods, in the case where the number of generate data grows significantly, compared to the dimension of the



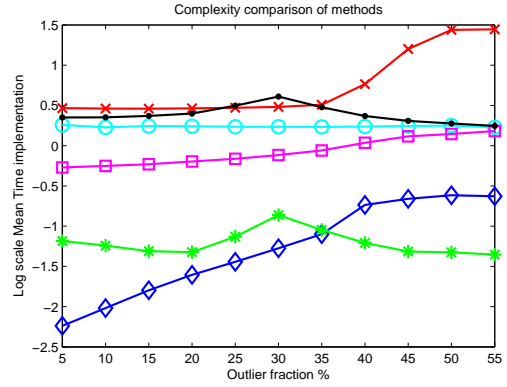
(a) Dimension of the unknown vector $m = 50$.



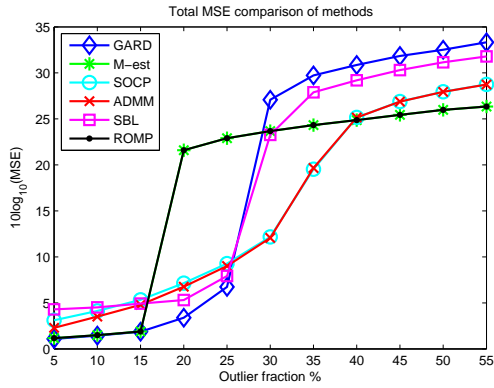
(b) Dimension of the unknown vector $m = 50$.



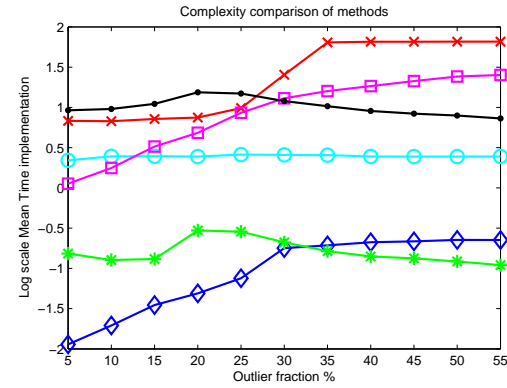
(c) Dimension of the unknown vector $m = 100$.



(d) Dimension of the unknown vector $m = 100$.



(e) Dimension of the unknown vector $m = 170$.



(f) Dimension of the unknown vector $m = 170$.

Figure 1: Figures (a), (c), (e): The attained MSE versus the outlier fraction, for various dimensions of the unknown vector θ_0 . In all cases, $n = 600$ observations were used. Figures (b), (d), (f): Log-scale of Mean Time versus the fraction of outliers.

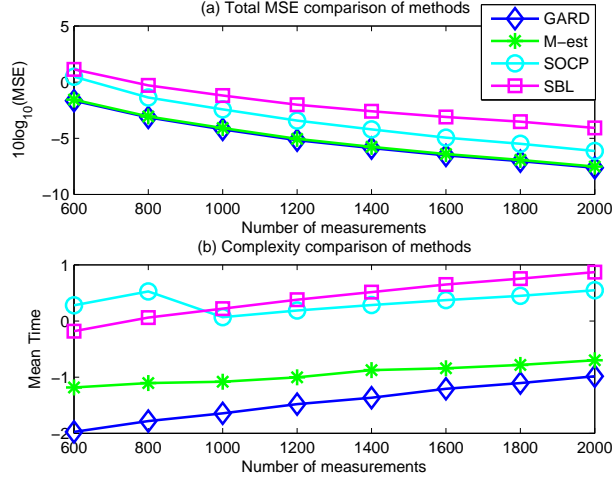


Figure 2: Large scale complexity test for dimension of the unknown vector set at $m = 100$. While varying the number of measurements, the MSE (a) and the Mean time until convergence (b), is shown for each method. It is clear that GARD attains the lowest MSE, whilst being the most efficient.

unknown vector θ_0 . Comparison is performed for all methods except from ADMM and ROMP. As presented in table 1, the ADMM algorithm does not handle efficiently large numbers of samples. On the other hand, although ROMP performs exactly as M-est, that comes at a higher computational cost, therefore seemed impractical to put it to test. Once again, equation (5) has been used to generate our data. The dimension of θ_0 is set at $m = 100$, the density of the outlier noise is 10% with values ± 25 spread uniformly over n coordinates and finally the inlier noise vector has elements drawn from the standard Gaussian distribution, with $\sigma = 1$ and inlier noise bound ϵ_0 , assumed to be known. For each number of measurements n , 100 independent experiments have been performed, while varying the inlier and outlier noise; results have been averaged.

In figure 2 (b), we have evaluated the mean implementation time (in logarithmic scale for precision) for each method, while at the same time the total MSE is measured, for each varying size of n . It is clear, that even for significantly large n , i.e., large number of measurements, GARD excels. Whilst attaining the lowest MSE, the mean time until convergence, is the lowest.

5.3 Support recovery test

This section attends to bridge the gap, between the theoretical results properties in section 4 and the experimental performance of GARD. The results of section 5.1, showcase the performance of GARD. However, it would be incorrect to conclude that the support of the sparse outlier vector is correctly recovered, in cases where the algorithm attains a low MSE, a matter that we would like to address here. Although, the recovery of the sparse outlier support is desirable, since it guarantees the smallest MSE possible, it should be noted that GARD performs well (with respect to the MSE), even in cases where the recovery of the support is not exact; e.g., one of the most common cases is to identify a few extra indices (that do not belong to the support of \mathbf{u}_0) as outlying elements.

For all support recovery tests, we have set the dimension of the unknown vector θ_0 , at $m = 100$ and corrupted the original data with outliers in $s < n$ indices, uniformly sampled over $n = 600$ measurements. Also, for each fraction of outliers, i.e., $(s/n) \cdot 100\%$, we performed 10000 Monte Carlo runs.

Let S_k denote the support set of the sparse estimate \mathbf{u}_* and S the support set of the sparse outlier vector \mathbf{u}_0 . The green line corresponds to the percentage of correct indices the proposed scheme has recovered, i.e., indices $i \in S_k \subseteq S$, while the orange line corresponds to the extra indices that the method has identified as outliers, i.e., indices $j \in J \setminus S$. In parallel, since the smallest principal angle cannot be computed directly, we have tracked the bound of $c > \delta_s$ for evaluation of the theoretical results proposed in section 4.2. The vertical line, corresponds to the largest outlier fraction, that the proposed scheme succeeds in recovering the sparse outlier vector support, one to one.

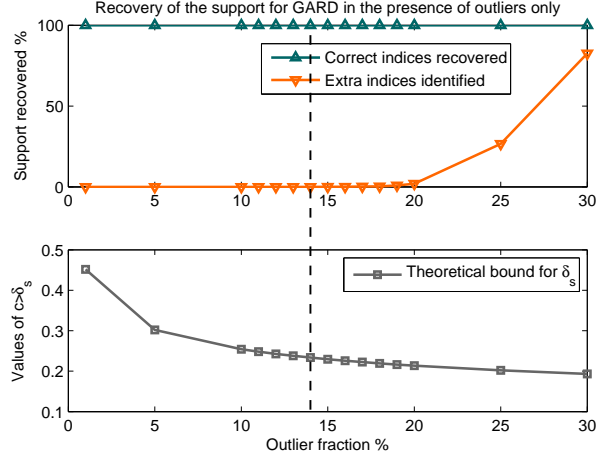


Figure 3: Recovery of the support and relation to the bound of δ_s , for the noiseless case. For outlier fraction of less than 14%, the bound for δ_s (24), is guaranteed, hence the recovery is exact.

5.3.1 The presence of outliers only

The scenario in which our original data is corrupted with outlier values only, is treated separately. Our data are generated via equation (5), for $\boldsymbol{\eta} = \mathbf{0}$ and outlier values⁷ ± 25 , in s indices, uniformly sampled over n coordinates. In figure 3, the recovery of the exact support versus fraction of outliers is demonstrated. It is clear that for fraction of less than 14%, the bound for δ_s as Theorem 3 proposes, is guaranteed, thus the recovery of the support is exact and also the approximation of $\boldsymbol{\theta}_0$ is of zero error. Noteworthy is also the fact that the approximation error is also zero, in cases where only a few extra indices that belong to S^c , are imported into the support set S_k .

5.3.2 The presence of both inlier and outlier noise

In the current section, we have worked towards the empirical validation of (25) and (29), where two separate tests have been performed. Equation (5), has been also employed here to generate our input data.

In the first test, we have fixed the maximum bound for the norm of the inlier noise vector at $\epsilon_0 = 28$, while we have altered the fraction of outliers. In order to achieve this, we have used Matlab's random generator for the Gaussian distribution with standard deviation depending on ϵ_0 , while we have cut off the largest elements (in the absolute sense) when it was required, so that the norm of the inlier noise vector always remains bounded by ϵ_0 . Also, recall on remark 11, that the minimum element of the absolute value of the outlier vector should be greater than $(2 + \sqrt{6})\epsilon_0$, in order (25) to be valid. Thus, outlier values have been set at ± 150 , while the values of $\mathbf{y}_0 = \mathbf{X}\boldsymbol{\theta}_0$, range at 170 – 180.

In figure 4, we have plotted the recovery of the support for GARD and its relation to the bound c of the smallest principal angle δ_s , for each outlier fraction. As one could observe, for fraction of outliers less than 13%, the bound for δ_s as Theorem 4 proposes is guaranteed, thus the recovery of the support is exact. In parallel, we have computed the MSE between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_*$ and tracked the relation to the theoretical bound⁸ of (29).

In the second test, the ability of GARD to deal with heavy loads of noise, is demonstrated. The outlier values were set at ± 150 and the bound of ϵ_0 was increased, so that the inlier noise corresponds to noise of 20 dB. In such a case, the bounds established in (25) and (29) are violated, however GARD manages to do well. In figure 5, the recovery of the support versus the outlier fraction is demonstrated. We conclude, that although the method does not succeed to recover the sparse outlier support 100%, the MSE is relatively low, at least for low fraction of outliers, i.e., below 10%. It should be noted that the MSE value close to 5 is not high, compared to the MSE measured in figure 4, which was close to 1.

⁷In the noiseless case, arbitrarily small values, are treated as outliers; thus the performance of GARD is not affected by a particular selection of those values.

⁸Since the MSE is a squared norm between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_*$, the bound is the squared right hand side of (29).

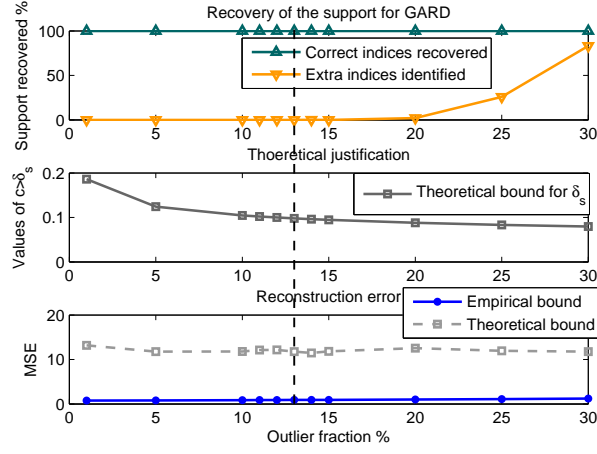


Figure 4: Recovery of the support and relation to the bound of δ_s , for the case inlier and outlier noise coexist. For outlier fraction of less than 13%, the bound for δ_s (24), is guaranteed, hence the recovery of the support is exact, while the MSE computed is valid under the bound that inequality (29) suggests.

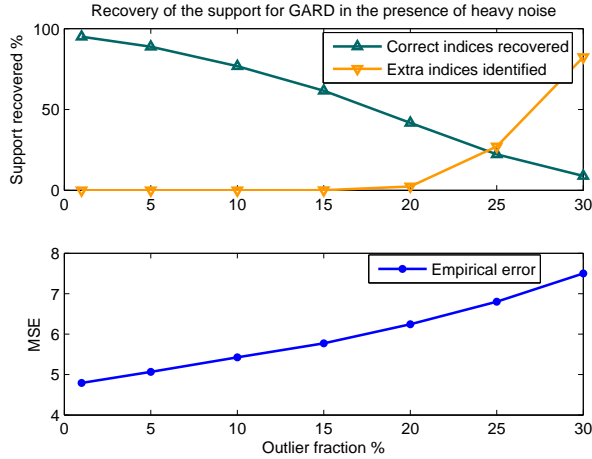
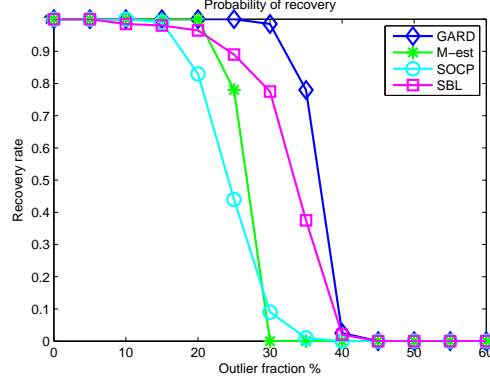
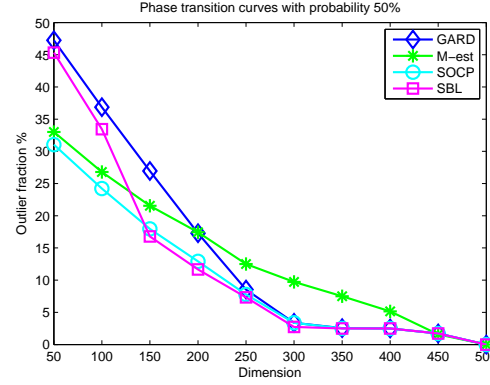


Figure 5: Recovery of the support for GARD, in the case where outlier and heavy inlier noise of approx. 20 dB coexist. Although, the support is not entirely recovered, the MSE is relatively low.



(a) Probability of recovery for dimension of the unknown vector $m = 100$.



(b) Phase Transition curves.

Figure 6: (a) Probability of recovery while varying the fraction of outliers. (b) Transition from success to failure of each method with probability 50%.

5.4 Phase transition curves

In the final experiment, we have tested whether each method succeeds to reach an estimate or not. In particular, we computed the probability for which each method succeeds in recovering the solution. Since ADMM and ROMP have a higher computational cost, as well as similar to other methods performance, we limited our efforts to test the other three methods, i.e., GARD, SOCP, M-est and SBL.

Figure 6 (a) shows the probability of recovery for each method tested, while varying the fraction of outliers. The dimension of the unknown vector θ_0 is $m = 100$. For each density of the sparse outlier vector, we have computed the probability over 200 Monte Carlo runs. For each method, we have assumed that the solution is found, if $\|\theta_* - \theta_0\|_2 / \|\theta_0\|_2 \leq 0.03$. The major result, is that for fraction of outliers under 25%, GARD succeeds in recovering the solution, with probability 1. For (M-est), the percentage is below 20%, while for the rest ℓ_1 minimization methods the percentage is even lower. For SBL, the probability to recover the solution is not guaranteed, even for the lowest fractions of outliers.

Figure 6 (b), shows the phase transition curves for each method. For each dimension of the unknown vector θ_0 , we have computed the fraction of outliers for which the method transits from success to failure with probability 50%. Experiments were carried over 200 Monte Carlo runs. Once again, we have assumed that the solution is found, using the criterion as in Figure 6 (a). We can observe that for each fixed dimension of the unknown vector the probability for each method to recover the solution (always within a given tolerance), increases for fraction of outliers below each phase transition curve, while the probability decreases as we move above the phase transition curves. Here, it is clear that up to $m = 200$, GARD succeeds to recover the solution with the highest probability from all the rest of the methods. However, for greater dimension of the unknown, the measurement dimension (here $n = 600$) seems pretty “poor” to allow GARD to preserve the performance (in the sense that more data is required), although it does not drop below the ℓ_1 minimization techniques.

Algorithm	Test-A	Test-B	Test-C	Test-D
GARD	0.1772	0.0180	0.0586	0.690
M-est	0.2248	0.2859	1.844e+06	0.704
SOCP	0.4990	0.3502	5.852e+05	1.011
SBL	0.9859	58.3489	2.165e+06	1.292
ROMP	0.2248	0.2859	1.844e+06	0.704

Table 2: Computed MSE, for various experiments. In tests A,B and C, the noise is drawn from the Heavy-tailed distribution alpha-stable of Levy distribution. In test D, noise consists of a sum of two vectors, drawn from 2 independent Gaussian distributions with different variance, plus an outlier noise vector of impulsive noise.

5.5 Experiments with general noise forms

In the current section we performed a set of more realistic experiments for the methods described and measured the MSE over an average of 100 Monte-Carlo runs. Equation (2), describes our model, where we produced ($n = 600$) measurements corrupted with different types of noise and measured the MSE. The dimension of the unknown vector θ_0 is $m = 100$. For all tests, the ADMM was excluded from the last set of experiments, since the method proved weak to handle different orders of noise values, thus failed to converge for all tests.

- **Tests A, B and C.** The noise vector is drawn from the Lévy alpha-stable distribution, $\mathcal{S}(\alpha, \beta, \gamma, \delta)$, with pdf expressed in closed form only for special cases of the parameters defined. The distribution's parameters β and δ that control symmetry, were set to zero (results to a symmetric distribution without skewness) for all three experiments. For A, the distribution's parameters were set to $\alpha = 0.45$ and $\gamma = 0.3$; the parameters for each method were set to $\epsilon_0 = 3$ for GARD and SOCP, $\sigma = 1.2$ for M-est and ROMP, while the hyperparameters for SBL were initialized to 10^{-4} . In table 2, it can be seen that almost all methods perform quite well (MSE is low), with GARD appearing to perform better. For test B, $\alpha = 0.4$, $\gamma = 0.1$; for GARD $\epsilon_0 = 3$, for SOCP $\epsilon_0 = 2$, for M-est and ROMP $\sigma = 1$, while for SBL the hyperparameters were initialized at random (Gaussian) with variance equal to 10^{-5} , although fails to converge, for all values of the parameters tested. Once again, it can be readily seen that GARD attains the lowest MSE. Finally, for experiment C, $\alpha = 0.3$, $\gamma = 0.1$, resulting to more large values of noise; for GARD $\epsilon_0 = 3$, for SOCP $\epsilon_0 = 2$, for M-est and ROMP $\sigma = 1$, while for SBL the hyperparameters were initialized at random (Gaussian) with variance equal to 10^{-6} . In table 2, the MSE for GARD is significant lower than tests A and B, which means that the method identifies correctly the outlier values, regardless how large those are; the rest of the methods fail to provide a descent estimate for the unknown vector.
- **Test D.** The noise consists of a sum of two vectors, drawn from 2 independent Gaussian distributions $\mathcal{N}(0, 0.6^2)$ and $\mathcal{N}(0, 0.8^2)$, plus an outlier noise vector of 10% density (indices chosen uniformly at each repetition) with values ± 25 . The parameters required for each method are: the default tuning parameter for both M-est and ROMP; for GARD and SOCP $\max\{\epsilon_1, \epsilon_2\}$ is required, where ϵ_1, ϵ_2 are the bounds of each inlier noise vector, while for SBL an initialization at random with variance of 10^{-6} was performed. The model of the noise is now more complicated, hence the problem harder to solve for all methods. Once again, it is clear that GARD succeeds in handling this mixed type of noise too.

6 Conclusions

A novel algorithmic scheme, i.e., GARD, for robust linear regression has been developed. GARD alternates between an OMP selection step, which identifies the outliers, and a Least-Squares estimator, that attempts to fit the data. Several properties regarding convergence, error bounds and uniqueness of the solution have been derived. Furthermore, more theoretical results concerning the stability of the method and the recovery of the outliers' support have been extracted.

The proposed scheme has been compared with other well established techniques through extensive simulations. The experiments suggest that GARD has an overall tolerance in outliers compared to its competitors. Moreover, it attains the lowest error for the estimation of the unknown vector, along with M-est and ROMP; moreover, GARD attains similar MSE at lower complexity.

Finally, the experiments verify that our greedy-based GARD algorithm outperforms the ℓ_1 norm-based schemes, for low sparsity levels; since in practical applications outliers are expected to be just a few, greedy-based techniques seem to be the right choice.

A Proof of Theorem 3

Since matrix \mathbf{X} is assumed to be full rank, according to the analysis presented in section 4.2, equation $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}_0 + \mathbf{u}_0$ could be transformed into equation (23). Thus, the principal angle defined in (20), is now involved.

Rather than delving into the main arguments of the proof, we need to establish the following propositions.

Proposition 1. *Let \mathbf{Q} be the orthonormal matrix of the reduced QR decomposition of the full rank matrix \mathbf{X} and δ_s the principal angle between the subspace spanned by $\text{span}(\mathbf{Q})$ and the subspace spanned by all the s -dimensional outlier subspaces. Then*

$$\|\mathbf{Q}^T \mathbf{v}\|_2 \leq \delta_s \|\mathbf{v}\|_2 \quad (30)$$

holds for every vector $\mathbf{v} \in \mathbf{R}^n$ with $\|\mathbf{v}\|_0 \leq s$.

Proof. The proof is straightforward by the definition of δ_s and equation (21):

$$\begin{aligned} \|\mathbf{Q}^T \mathbf{v}\|_2^2 &= |\langle \mathbf{v}, \mathbf{Q}\mathbf{Q}^T \mathbf{v} \rangle| \leq \delta_s \|\mathbf{v}\|_2 \|\mathbf{Q}\mathbf{Q}^T \mathbf{v}\|_2 \\ &\leq \delta_s \|\mathbf{v}\|_2 \|\mathbf{Q}\|_2 \|\mathbf{Q}^T \mathbf{v}\|_2 = \delta_s \|\mathbf{v}\|_2 \|\mathbf{Q}^T \mathbf{v}\|_2, \end{aligned}$$

which leads to (30). \square

Lemma 3. *Let the assumptions of Proposition 1 be satisfied and S be any non-empty subset of $J = 1:n$ with cardinality $|S| = k \leq s < n$. Then*

$$\|\mathbf{Q}^T \mathbf{I}_S\|_2 \leq \delta_s \quad (31)$$

holds for every such set S .

Proof. Let $\mathbf{v} \neq \mathbf{0}$ be a vector of \mathbf{R}^k , $k \leq s$. It is clear that $\mathbf{I}_S \mathbf{v} = \mathbf{v} \in \mathbf{R}^n$, with $\|\mathbf{v}\|_0 \leq s$ and $\|\mathbf{v}\|_2 = \|\mathbf{v}\|_2$. Thus, for all $\mathbf{v} \in \mathbf{R}^k$ we have

$$\|\mathbf{Q}^T \mathbf{I}_S \mathbf{v}\|_2 = \|\mathbf{Q}^T \mathbf{v}\|_2 \leq \delta_s \|\mathbf{v}\|_2,$$

due to Proposition 1. The result follows from the definition of the matrix 2-norm. \square

The importance of Lemma 3 is twofold. First of all, it is a bound on the 2-norm of the matrix $\mathbf{Q}^T \mathbf{I}_S$. Moreover, since $\|\mathbf{I}_S^T \mathbf{Q}\mathbf{Q}^T \mathbf{I}_S\|_2 = \|\mathbf{Q}^T \mathbf{I}_S\|_2^2$ and assuming that (24) holds, we have that

$$\|\mathbf{I}_S^T \mathbf{Q}\mathbf{Q}^T \mathbf{I}_S\|_2 \leq \delta_s^2 < 1/2, \quad (32)$$

which leads to the fact that matrix $\mathbf{M}_k = \mathbf{I}_k - \mathbf{I}_S^T \mathbf{Q}\mathbf{Q}^T \mathbf{I}_S$ is invertible for all S with $|S| = k \leq s$ and

$$\|\mathbf{M}_k^{-1}\|_2 \leq (1 - \|\mathbf{I}_S^T \mathbf{Q}\mathbf{Q}^T \mathbf{I}_S\|_2)^{-1} < 2, \quad (33)$$

due to a very popular Proposition of linear algebra.

Lemma 4. *Let the assumptions of Lemma 3 be satisfied. Then*

$$\|\mathbf{I}_S^T \mathbf{Q}\mathbf{Q}^T \mathbf{v}\|_2 \leq \delta_s^2 \|\mathbf{v}\|_2 \quad (34)$$

holds for every vector $\mathbf{v} \in \mathbf{R}^n$, with $\|\mathbf{v}\|_0 \leq s$.

Proof. The tricky part of the proof is that the s -sparse vector $\mathbf{v} \in \mathbf{R}^n$, might not necessarily share the same support set S . Let S' the support set of vector \mathbf{v} , with $|S'| = k \leq s$. Thus, using $\mathbf{v}_{S'} \in \mathbf{R}^k$ to denote the non-sparse vector (also notice that $\|\mathbf{v}\|_2 = \|\mathbf{v}_{S'}\|_2$), we have $\mathbf{v} = \mathbf{I}_{S'} \mathbf{v}_{S'}$. Hence, due to the sub-multiplicative property of the matrix 2-norm, we have

$$\begin{aligned} \|\mathbf{I}_S^T \mathbf{Q}\mathbf{Q}^T \mathbf{v}\|_2 &= \|\mathbf{I}_S^T \mathbf{Q}\mathbf{Q}^T \mathbf{I}_{S'} \mathbf{v}_{S'}\|_2 \\ &\leq \|\mathbf{I}_S^T \mathbf{Q}\|_2 \|\mathbf{Q}^T \mathbf{I}_{S'}\|_2 \|\mathbf{v}_{S'}\|_2 \leq \delta_s^2 \|\mathbf{v}\|_2, \end{aligned}$$

where we have used $\|\mathbf{I}_S^T \mathbf{Q}\|_2 = \|\mathbf{Q}^T \mathbf{I}_S\|_2$ and both the results of Proposition 1 and Lemma 3. \square

Proof of the Main Theorem

Proof. Let $S = \text{supp}(\mathbf{u}_0)$. At the initial step of (GARD), the Least Squares solution over the active columns of matrix $\Phi = [\mathbf{Q} \mathbf{I}_n]$ is computed, i.e., columns vectors of matrix \mathbf{Q} . The initial residual is $\mathbf{r}^{(0)} = \mathbf{y} - \mathbf{Q}\mathbf{w}_*^{(0)}$. At this point, we could express the residual in terms of the projection matrix $\mathbf{P}_\mathbf{Q}$ onto the range of matrix \mathbf{Q} ⁹. Thus, $\mathbf{P}_\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T$ and the residual could be written as

$$\mathbf{r}^{(0)} = (\mathbf{I}_n - \mathbf{P}_\mathbf{Q})\mathbf{y} = (\mathbf{I}_n - \mathbf{P}_\mathbf{Q})\mathbf{u}_0,$$

as $\mathbf{y} = \mathbf{Q}\mathbf{w}_0 + \mathbf{u}_0$, $\mathbf{I}_n - \mathbf{P}_\mathbf{Q}$ is the projector for the subspace complementary to that of $\mathcal{R}(\mathbf{Q})$ and $(\mathbf{I}_n - \mathbf{P}_\mathbf{Q})\mathbf{Q}\mathbf{w}_0 = \mathbf{0}$.

At the first step, in order to ensure a selection from the correct support S , we impose that

$$|r^{(0)}(i)| > |r^{(0)}(j)|, \quad \forall i \in S \text{ and } j \in S^c. \quad (35)$$

The basic concept of the proof is to obtain lower and upper bounds for the left and right part of equation (35). Employing Lemma 4, the left part is bounded below by

$$\begin{aligned} |r^{(0)}(i)| &= |\langle \mathbf{r}^{(0)}, \mathbf{e}_i \rangle| = |\langle \mathbf{u}_0 - \mathbf{Q}\mathbf{Q}^T \mathbf{u}_0, \mathbf{e}_i \rangle| \geq \\ &\geq |u_i| - |\langle \mathbf{Q}\mathbf{Q}^T \mathbf{u}_0, \mathbf{e}_i \rangle| = |u_i| - |\mathbf{e}_i^T \mathbf{Q}\mathbf{Q}^T \mathbf{u}_0| \\ &\geq \min_{l \in S} |u_l| - \delta_s^2 \|\mathbf{u}_0\|_2, \end{aligned} \quad (36)$$

where u_i are the elements of \mathbf{u}_0 .

Following similar steps, the right part is upper bounded by

$$\begin{aligned} |r^{(0)}(j)| &= |\langle \mathbf{r}^{(0)}, \mathbf{e}_j \rangle| = |\langle \mathbf{u}_0 - \mathbf{Q}\mathbf{Q}^T \mathbf{u}_0, \mathbf{e}_j \rangle| = \\ &= |\mathbf{e}_j^T \mathbf{Q}\mathbf{Q}^T \mathbf{u}_0| \leq \delta_s^2 \|\mathbf{u}_0\|_2, \end{aligned} \quad (37)$$

using that $\langle \mathbf{u}_0, \mathbf{e}_j \rangle = 0$, since $j \in S^c$.

Hence, if we impose

$$\min_{l \in S} |u_l| - \delta_s^2 \|\mathbf{u}_0\|_2 > \delta_s^2 \|\mathbf{u}_0\|_2,$$

condition (35) is guaranteed and one of the correct columns, i.e., i_1 , is bound to be selected at the first step (note, that it is not guaranteed that the largest outlier value will be selected first).

Considering $S_1 = \{j_1\} \subset S$, the matrix of active columns is augmented, i.e., $\Phi_1 = [\mathbf{Q} \mathbf{e}_{j_1}]$ and the new residual is computed, requiring the inversion of

$$\Phi_1^T \Phi_1 = \begin{bmatrix} \mathbf{I}_m & \mathbf{Q}^T \mathbf{e}_{j_1} \\ \mathbf{e}_{j_1}^T \mathbf{Q} & 1 \end{bmatrix}.$$

Taking into account that \mathbf{I}_m is invertible and $\beta = 1 - \|\mathbf{Q}^T \mathbf{e}_{j_1}\|_2^2 > 1/2$ (inequality (32) for $|S| = 1$) and using the *Matrix Inversion Lemma* in block form, we obtain:

$$(\Phi_1^T \Phi_1)^{-1} = \begin{bmatrix} \mathbf{I}_m + \mathbf{Q}^T \mathbf{e}_{j_1} \mathbf{e}_{j_1}^T \mathbf{Q} / \beta & -\mathbf{Q}^T \mathbf{e}_{j_1} / \beta \\ -\mathbf{e}_{j_1}^T \mathbf{Q} / \beta & 1 / \beta \end{bmatrix}.$$

After a few lines of elementary algebra, we take

$$\begin{aligned} \Phi_1(\Phi_1^T \Phi_1)^{-1} \Phi_1^T &= \mathbf{Q}\mathbf{Q}^T + \mathbf{Q}\mathbf{Q}^T \mathbf{e}_{j_1} \mathbf{e}_{j_1}^T \mathbf{Q}\mathbf{Q}^T / \beta - \\ &- \mathbf{e}_{j_1} \mathbf{e}_{j_1}^T \mathbf{Q}\mathbf{Q}^T / \beta - \mathbf{Q}\mathbf{Q}^T \mathbf{e}_{j_1} \mathbf{e}_{j_1}^T / \beta + \mathbf{e}_{j_1} \mathbf{e}_{j_1}^T / \beta. \end{aligned}$$

Hence, the new residual $\mathbf{r}^{(1)} = \mathbf{y} - \Phi_1(\Phi_1^T \Phi_1)^{-1} \Phi_1^T \mathbf{y}$, can be recast as

$$\begin{aligned} \mathbf{r}^{(1)} &= (\mathbf{I}_n - \mathbf{Q}\mathbf{Q}^T - \mathbf{Q}\mathbf{Q}^T \mathbf{e}_{j_1} \mathbf{e}_{j_1}^T \mathbf{Q}\mathbf{Q}^T / \beta + \mathbf{e}_{j_1} \mathbf{e}_{j_1}^T \mathbf{Q}\mathbf{Q}^T / \beta \\ &+ \mathbf{Q}\mathbf{Q}^T \mathbf{e}_{j_1} \mathbf{e}_{j_1}^T / \beta - \mathbf{e}_{j_1} \mathbf{e}_{j_1}^T) \mathbf{u}_0. \end{aligned} \quad (38)$$

Relation (38) could be simplified using the decomposition for the outlier vector, i.e., $\mathbf{u}_0 = u_{j_1} \mathbf{e}_{j_1} + \mathbf{F}_{S \setminus S_1}(\mathbf{u}_0)$, where the second part is the vector which has the same elements as \mathbf{u}_0 over the set $S \setminus S_1$, besides the j_1 -th

⁹Take into account that \mathbf{Q} is orthonormal.

coordinate which is equal to zero. Obviously, this is a vector, $s-1$ sparse at most and its support is a subset of the support of \mathbf{u}_0 . Thus, we have:

$$\begin{aligned}\mathbf{r}^{(1)} &= (\mathbf{I}_n - \mathbf{Q}\mathbf{Q}^T - \mathbf{Q}\mathbf{Q}^T \mathbf{e}_{j_1} \mathbf{e}_{j_1}^T \mathbf{Q}\mathbf{Q}^T / \beta + \\ &\quad + \mathbf{e}_{j_1} \mathbf{e}_{j_1}^T \mathbf{Q}\mathbf{Q}^T / \beta) F_{S \setminus S_1}(\mathbf{u}_0) = \\ &= \mathbf{v}_1 - \mathbf{Q}\mathbf{Q}^T \mathbf{v}_1,\end{aligned}\tag{39}$$

where $\mathbf{v}_1 = F_{S \setminus S_1}(\mathbf{u}_0) + \gamma_1 \mathbf{e}_{j_1}$, $\gamma_1 = \mathbf{e}_{j_1}^T \mathbf{Q}\mathbf{Q}^T F_{S \setminus S_1}(\mathbf{u}_0) / \beta$.

At this point, we should note that $\text{supp}(\mathbf{v}_1) = \text{supp}(\mathbf{u}_0) = S$, while $\gamma_1 \neq u_{j_1}$. Following a similar rational, for the next step, we impose $|r^{(1)}(i)| > |r^{(1)}(j)|$ for all $i \in S \setminus S_1$ and $j \in S^c$. Hence, using lower and upper bounds leads to

$$\begin{aligned}|r^{(1)}(i)| &= |\langle \mathbf{r}^{(1)}, \mathbf{e}_i \rangle| = |\langle \mathbf{v}_1 - \mathbf{Q}\mathbf{Q}^T \mathbf{v}_1, \mathbf{e}_i \rangle| \geq \\ &\geq |u_i| - |\mathbf{e}_i^T \mathbf{Q}\mathbf{Q}^T \mathbf{v}_1| \geq \min_{l \in S} |u_l| - \delta_s^2 \|\mathbf{v}_1\|_2,\end{aligned}\tag{40}$$

where we used that $\langle \mathbf{e}_{i_1}, \mathbf{e}_i \rangle = 0$ for $i \in S \setminus S_1$, and

$$\begin{aligned}|r^{(1)}(j)| &= |\langle \mathbf{r}^{(1)}, \mathbf{e}_j \rangle| = |\langle \mathbf{v}_1 - \mathbf{Q}\mathbf{Q}^T \mathbf{v}_1, \mathbf{e}_j \rangle| = \\ &= |\mathbf{e}_j^T \mathbf{Q}\mathbf{Q}^T \mathbf{v}_1| \leq \delta_s^2 \|\mathbf{v}_1\|_2,\end{aligned}\tag{41}$$

where we exploited the relationship $\langle \mathbf{v}_1, \mathbf{e}_j \rangle = 0$, for every $j \in S^c$, as well as lemma 4.

Imposing $\min_{l \in S} |u_l| - \delta_s^2 \|\mathbf{v}_1\|_2 > \delta_s^2 \|\mathbf{v}_1\|_2$, leads equivalently to

$$\delta_s < \sqrt{\frac{\min_{l \in S} |u_l|}{2 \|\mathbf{v}_1\|_2}}.\tag{42}$$

Although (42), seems inadequate, we will show indeed that this is a condition which always holds true, provided (24) is satisfied. One needs to prove that $\|\mathbf{u}_0\|_2 > \|\mathbf{v}_1\|_2$, which is equivalent to showing that $|\gamma_1| < |u_{j_1}|$, using the aforementioned decompositions of \mathbf{u}_0 , \mathbf{v}_1 and the Pythagorean Theorem. Thus, we have that

$$\begin{aligned}|\gamma_1| &= \frac{|\langle \mathbf{e}_{j_1}, \mathbf{Q}\mathbf{Q}^T F_{S \setminus S_1}(\mathbf{u}_0) \rangle|}{|\beta|} \leq 2\delta_s^2 \|F_{S \setminus S_1}(\mathbf{u}_0)\|_2 \\ &< \min_{j \in S} |u_j| \leq |u_{j_1}|,\end{aligned}\tag{43}$$

due to $\beta > 1/2$, the definition of the principal angle (19), inequality (24) and $\|F_{S \setminus S_1}(\mathbf{u}_0)\|_2 < \|\mathbf{u}_0\|_2$, for any non-empty set S_1 .

At the k -step $S_k = \{j_1, j_2, \dots, j_k\} \subset S$ and the matrix that corresponds to the set of active columns is $\Phi_{\mathbf{k}} = [\mathbf{Q} \mathbf{I}_{S_k}]$. Using again the *Matrix Inversion Lemma* for the inversion of $\Phi_{\mathbf{k}}^T \Phi_{\mathbf{k}}$, the new residual is given as follows:

$$\begin{aligned}\mathbf{r}^{(k)} &= (\mathbf{I}_n - \mathbf{Q}\mathbf{Q}^T - \mathbf{Q}\mathbf{Q}^T \mathbf{I}_{S_k} \mathbf{M}_k^{-1} \mathbf{I}_{S_k}^T \mathbf{Q}\mathbf{Q}^T + \\ &\quad + \mathbf{I}_{S_k} \mathbf{M}_k^{-1} \mathbf{I}_{S_k}^T \mathbf{Q}\mathbf{Q}^T) F_{S \setminus S_k}(\mathbf{u}_0) \\ &= \mathbf{v}_k - \mathbf{Q}\mathbf{Q}^T \mathbf{v}_k,\end{aligned}\tag{44}$$

where we used the identities

$$\mathbf{M}_k = \mathbf{I}_k - \mathbf{I}_{S_k}^T \mathbf{Q}\mathbf{Q}^T \mathbf{I}_{S_k},\tag{45}$$

$$\mathbf{u}_0 = F_{S_k}(\mathbf{u}_0) + F_{S \setminus S_k}(\mathbf{u}_0),\tag{46}$$

$$\mathbf{v}_k = F_{S \setminus S_k}(\mathbf{u}_0) + \mathbf{I}_{S_k} \mathbf{M}_k^{-1} \mathbf{I}_{S_k}^T \mathbf{Q}\mathbf{Q}^T F_{S \setminus S_k}(\mathbf{u}_0),\tag{47}$$

It is not hard to verify that $\text{supp}(\mathbf{v}_k) = \text{supp}(\mathbf{u}_0) = S$ still holds true. For a correct outlier index selection from the set S , at $k+1$ step, one needs to impose $|r^{(k)}(i)| > |r^{(k)}(j)|$ for all $i \in S \setminus S_k$ and $j \in S^c$. Using lower and upper bounds on the inner products, one obtains relations similar to (40), (41) with \mathbf{v}_k instead of \mathbf{v}_1 , which leads to

$$\delta_s < \sqrt{\frac{\min_{l \in S} |u_l|}{2 \|\mathbf{v}_k\|_2}}.\tag{48}$$

The proof ends, by showing that the last bound is looser than that of inequality (24), simply by proving that $\|\mathbf{v}_k\|_2 < \|\mathbf{u}_0\|_2$ for all $k = 1, 2, \dots, s-1$.

Using the decompositions of these vectors (46), (47) and the Pythagorean Theorem, it suffices to show that $\|\mathbf{M}_k^{-1} \mathbf{I}_{S_k}^T \mathbf{Q} \mathbf{Q}^T F_{S \setminus S_k}(\mathbf{u}_0)\|_2 < \|F_{S_k}(\mathbf{u}_0)\|_2$, which follows from the fact that

$$\begin{aligned} & \|\mathbf{M}_k^{-1} \mathbf{I}_{S_k}^T \mathbf{Q} \mathbf{Q}^T F_{S \setminus S_k}(\mathbf{u}_0)\|_2 \\ & \leq \|\mathbf{M}_k^{-1}\|_2 \|\mathbf{I}_{S_k}^T \mathbf{Q} \mathbf{Q}^T F_{S \setminus S_k}(\mathbf{u}_0)\|_2 \\ & < \min_{l \in S} |u_l| \leq \|F_{S_k}(\mathbf{u}_0)\|_2, \end{aligned} \tag{49}$$

where we employed the sub-multiplicative property of the matrix 2-norm, inequality (33), Lemma 4 and (24).

Finally, at step $k+1 = s$ it is guaranteed that the correct support is recovered and thus the linear subspace, onto which the measurement vector \mathbf{y} lies, is build. In turn, this results to a Least Squares solution for GARD of zero error. \square

B Proof of Theorem 4

Since, theorem 4 is the generalization of 3, some intermediate results regarding the proof presented in Appendix A, will also be used here. On the other hand, we will try to avoid the technical parts with shared similarities.

Proof. Due to the existence and uniqueness of the QR decomposition, the analysis is based on equation (17).

Since initially GARD performs a Least Squares step, where the columns that participate in the representation are only those of matrix \mathbf{X} , the residual is also $\mathbf{r}^{(0)} = (\mathbf{I}_n - \mathbf{Q} \mathbf{Q}^T) \mathbf{y}$. Thus, taking into account (17), we have the following expression for the initial residual:

$$\mathbf{r}^{(0)} = \mathbf{u}_0 + \boldsymbol{\eta} - \mathbf{Q} \mathbf{Q}^T \mathbf{u}_0 - \mathbf{Q} \mathbf{Q}^T \boldsymbol{\eta},$$

where the extra terms are due to the noise vector $\boldsymbol{\eta}$.

Once again, we should impose (35), according to (36) and (37). Also, recall on Theorem 3, suggesting¹⁰ that $\delta_s < c \leq \sqrt{2}/2$. Thus, we have:

$$\begin{aligned} |r^{(0)}(i)| & \geq |u_i| - |\langle \mathbf{Q} \mathbf{Q}^T \mathbf{u}_0, \mathbf{e}_i \rangle| - |\langle \boldsymbol{\eta}, \mathbf{e}_i \rangle| - |\langle \mathbf{Q} \mathbf{Q}^T \boldsymbol{\eta}, \mathbf{e}_i \rangle| \\ & \geq \min_{l \in S} |u_l| - \delta_s^2 \|\mathbf{u}_0\|_2 - \epsilon_0 - \epsilon_0 \delta_s \\ & > \min_{l \in S} |u_l| - \delta_s^2 \|\mathbf{u}_0\|_2 - \epsilon_0 - \frac{\epsilon_0}{\sqrt{2}} \\ & > \min_{l \in S} |u_l| - \delta_s^2 \|\mathbf{u}_0\|_2 - \epsilon_0 - \epsilon_0 \sqrt{\frac{3}{2}} \end{aligned}$$

and

$$\begin{aligned} |r^{(0)}(j)| & \leq \epsilon_0 + \delta_s^2 \|\mathbf{u}_0\|_2 + \epsilon_0 \delta_s \\ & < \epsilon_0 + \delta_s^2 \|\mathbf{u}_0\|_2 + \frac{\epsilon_0}{\sqrt{2}} \\ & < \epsilon_0 + \delta_s^2 \|\mathbf{u}_0\|_2 + \epsilon_0 \sqrt{\frac{3}{2}}, \end{aligned}$$

for $i \in S$ and $j \in S^c$, respectively. Thus, inequality (25) follows for the initial step. In the following, we proceed with the general selection step at $k+1$, as the first one is omitted, since it could be viewed as a special case of the general $k+1$ step. In the proof of Theorem 3, it was presented for comprehension reasons solely. It should also be noted, that the matrices augmented and inverted at each step, are those presented in the proof of Theorem 3. However, this is not the case for the solution and the residual, which is in our greatest interest.

¹⁰In the noiseless case, $\sqrt{2}/2$ was the upper bound for $c > \delta_s$, achieved only for 1-sparse outlier vectors. Thus, if δ_s exceeds this limit, GARD has little chance in recovering the correct support, even in the presence of outlier noise only, let alone as inlier noise coexists.

The condition in (25), guarantees, that at each selection step the support of our sparse estimate is a subset of the sparse outlier vector \mathbf{u}_0 , i.e., $S_k \subset S$ and the matrix that corresponds to the set of active columns is $\Phi_k = [\mathbf{Q} \mathbf{I}_{S_k}]$. Employing familiar techniques, we have an expression for the residual at the k step:

$$\mathbf{r}^{(k)} = \mathbf{v}_k + \boldsymbol{\eta}_k - \mathbf{Q}\mathbf{Q}^T \mathbf{v}_k - \mathbf{Q}\mathbf{Q}^T \boldsymbol{\eta}_k, \quad (50)$$

where \mathbf{v}_k is the vector defined in (47) and

$$\boldsymbol{\eta}_k = F_{J \setminus S_k}(\boldsymbol{\eta}) + \mathbf{I}_{S_k} \mathbf{M}_k^{-1} \mathbf{I}_{S_k}^T \mathbf{Q}\mathbf{Q}^T F_{J \setminus S_k}(\boldsymbol{\eta}). \quad (51)$$

In (51), it is clear that the only differences between $\boldsymbol{\eta}_k$ and $\boldsymbol{\eta}$, take place at the elements indexed $j_k \in S_k$, i.e, indices that GARD has selected as outliers. Moreover, for $J' = J \setminus S_k$ holds $J' \cap S_k = \emptyset$, i.e., the vector is decomposed into two disjoint subsets. At this point, prior to completing the proof, it is required to establish appropriate bounds for the inner products $|\langle \mathbf{e}_i, \mathbf{Q}\mathbf{Q}^T \boldsymbol{\eta}_k \rangle|$ and $|\langle \mathbf{e}_i, \boldsymbol{\eta}_k \rangle|$. Due to the Pythagorean Theorem, (31) and (33)

$$\begin{aligned} \|\boldsymbol{\eta}_k\|_2^2 &= \|F_{J \setminus S_k}(\boldsymbol{\eta})\|_2^2 + \|\mathbf{M}_k^{-1} \mathbf{I}_{S_k}^T \mathbf{Q}\mathbf{Q}^T F_{J \setminus S_k}(\boldsymbol{\eta})\|_2^2 \\ &\leq \epsilon_0^2 + 2\epsilon_0^2 = 3\epsilon_0^2. \end{aligned}$$

Hence,

$$|\mathbf{e}_i^T \mathbf{Q}\mathbf{Q}^T \boldsymbol{\eta}_k| \leq \delta_s \|\boldsymbol{\eta}_k\|_2 \leq \delta_s \sqrt{3}\epsilon_0 < \epsilon_0 \sqrt{\frac{3}{2}},$$

where we have also used the maximum bound, that $\delta_s < \sqrt{2}/2$. Also, for all $i \in J \setminus S_k$, holds $|\langle \mathbf{e}_i, \boldsymbol{\eta}_k \rangle| \leq |\langle \mathbf{e}_i, F_{J \setminus S_k}(\boldsymbol{\eta}_k) \rangle| \leq \epsilon_0$. Thus, adopting bounds to the absolute value of the inner products, we have

$$\begin{aligned} |r^{(k)}(i)| &\geq |u_i| - |\langle \mathbf{Q}\mathbf{Q}^T \mathbf{v}_k, \mathbf{e}_i \rangle| - \\ &\quad - |\langle \boldsymbol{\eta}_k, \mathbf{e}_i \rangle| - |\langle \mathbf{Q}\mathbf{Q}^T \boldsymbol{\eta}_k, \mathbf{e}_i \rangle| \geq \\ &\geq \min_{l \in S} |u_l| - \delta_s^2 \|\mathbf{v}_k\|_2 - \epsilon_0 - \epsilon_0 \sqrt{\frac{3}{2}} \end{aligned}$$

and

$$|r^{(k)}(j)| \leq \delta_s^2 \|\mathbf{v}_k\|_2 + \epsilon_0 + \epsilon_0 \sqrt{\frac{3}{2}},$$

for $i \in S \setminus S_k$ and $j \in S^c$, respectively. Thus, imposing $|r^{(k)}(i)| > |r^{(k)}(j)|$, leads to

$$\delta_s < \sqrt{\frac{\min_{l \in S} |u_l| - (2 + \sqrt{6})\epsilon_0}{2\|\mathbf{v}_k\|_2}},$$

which is satisfied, suppose (25) holds true. This holds true, due to the fact that $\|\mathbf{v}_k\|_2 < \|\mathbf{u}_o\|_2$ for all $k = 1, 2, \dots, s-1$ (read at the end of Appendix A for the proof). \square

References

- [1] W. J. Dixon *et al.*, “Analysis of extreme values,” *The Annals of Mathematical Statistics*, vol. 21, no. 4, pp. 488–506, 1950.
- [2] F. E. Grubbs, “Procedures for detecting outlying observations in samples,” *Technometrics*, vol. 11, no. 1, pp. 1–21, 1969.
- [3] D. M. Hawkins, *Identification of outliers*. Springer, 1980, vol. 11.
- [4] V. Barnett and T. Lewis, *Outliers in statistical data*. Wiley New York, 1994, vol. 3.
- [5] D. S. Moore and G. P. McCabe, *Introduction to the Practice of Statistics*. WH Freeman/Times Books/Henry Holt & Co, 1989.
- [6] S. Ramaswamy, R. Rastogi, and K. Shim, “Efficient algorithms for mining outliers from large data sets,” in *ACM SIGMOD Record*, vol. 29, no. 2. ACM, 2000, pp. 427–438.

- [7] J. Tang, Z. Chen, A. W.-C. Fu, and D. W. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2002, pp. 535–548.
- [8] Z. Chen, J. Tang, and A. W.-C. Fu, "Modeling and efficient mining of intentional knowledge of outliers," in *Database Engineering and Applications Symposium, 2003. Proceedings. Seventh International*. IEEE, 2003, pp. 44–53.
- [9] J. Tang, Z. Chen, A. W. Fu, and D. W. Cheung, "Capabilities of outlier detection schemes in large datasets, framework and methodologies," *Knowledge and Information Systems*, vol. 11, no. 1, pp. 45–84, 2007.
- [10] P. J. Huber, "The 1972 wald lecture robust statistics: A review," *The Annals of Mathematical Statistics*, pp. 1041–1067, 1972.
- [11] P. J. Rousseeuw and B. C. Van Zomeren, "Unmasking multivariate outliers and leverage points," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 633–639, 1990.
- [12] A. M. Leroy and P. J. Rousseeuw, "Robust regression and outlier detection," *J. Wiley&Sons, New York*, 1987.
- [13] P. J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*. John Wiley & Sons, 2005, vol. 589.
- [14] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [15] R. A. Maronna, R. D. Martin, and V. J. Yohai, *Robust statistics*. J. Wiley, 2006.
- [16] P. J. Huber, *Wiley Series in Probability and Mathematics Statistics*. Wiley Online Library, 1981.
- [17] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM journal on scientific computing*, vol. 20, no. 1, pp. 33–61, 1998.
- [18] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on pure and applied mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.
- [19] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
- [20] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse mri: The application of compressed sensing for rapid mr imaging," *Magnetic resonance in medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.
- [21] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *Information Theory, IEEE Transactions on*, vol. 52, no. 2, pp. 489–509, 2006.
- [22] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*. Academic Press, 2015.
- [23] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *Information Theory, IEEE Transactions on*, vol. 53, no. 12, pp. 4655–4666, 2007.
- [24] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *Constructive approximation*, vol. 13, no. 1, pp. 57–98, 1997.
- [25] S. Mallat, *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.
- [26] D. Needell and R. Vershynin, "Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 2, pp. 310–316, 2010.
- [27] D. Needell and J. A. Tropp, "Cosamp: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

- [28] E. J. Candes and T. Tao, “Decoding by linear programming,” *Information Theory, IEEE Transactions on*, vol. 51, no. 12, pp. 4203–4215, 2005.
- [29] R. J. Tibshirani *et al.*, “The lasso problem and uniqueness,” *Electronic Journal of Statistics*, vol. 7, pp. 1456–1490, 2013.
- [30] Y. Jin and B. D. Rao, “Algorithms for robust linear regression by exploiting the connection to sparse signal recovery,” in *Acoustics Speech and Signal Processing (ICASSP), International Conference on*. IEEE, 2010, pp. 3830–3833.
- [31] G. Mateos and G. B. Giannakis, “Robust nonparametric regression via sparsity control with application to load curve data cleansing,” *Signal Processing, IEEE Transactions on*, vol. 60, no. 4, pp. 1571–1584, 2012.
- [32] K. Mitra, A. Veeraraghavan, and R. Chellappa, “Robust rvm regression using sparse outlier model,” in *Computer Vision and Pattern Recognition (CVPR), Conference on*. IEEE, 2010, pp. 1887–1894.
- [33] S. A. Razavi, E. Ollila, and V. Koivunen, “Robust greedy algorithms for compressed sensing,” in *Signal Processing Conference (EUSIPCO), Proceedings of the 20th European*. IEEE, 2012, pp. 969–973.
- [34] J. A. Tropp, “Greed is good: Algorithmic results for sparse approximation,” *Information Theory, IEEE Transactions on*, vol. 50, no. 10, pp. 2231–2242, 2004.
- [35] S. Boyd, “Alternating direction method of multipliers,” in *Talk at NIPS Workshop on Optimization and Machine Learning*, 2011.
- [36] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.
- [37] D. W. Peaceman and H. H. Rachford, Jr, “The numerical solution of parabolic and elliptic differential equations,” *Journal of the Society for Industrial & Applied Mathematics*, vol. 3, no. 1, pp. 28–41, 1955.
- [38] J. Douglas, Jr, “On the numerical integration of $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = u$ by implicit methods,” *Journal of the Society for Industrial & Applied Mathematics*, vol. 3, no. 1, pp. 42–65, 1955.
- [39] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, “Applications of second-order cone programming,” *Linear algebra and its applications*, vol. 284, 1998.
- [40] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
- [41] M. E. Tipping, “Sparse bayesian learning and the relevance vector machine,” *The journal of machine learning research*, vol. 1, pp. 211–244, 2001.
- [42] D. P. Wipf and B. D. Rao, “Sparse bayesian learning for basis selection,” *Signal Processing, IEEE Transactions on*, vol. 52, no. 8, pp. 2153–2164, 2004.
- [43] B. L. Sturmfels and M. G. Christensen, “Comparison of orthogonal matching pursuit implementations,” in *Signal Processing Conference (EUSIPCO), Proceedings of the 20th European*. IEEE, 2012, pp. 220–224.
- [44] K. Mitra, A. Veeraraghavan, and R. Chellappa, “Analysis of sparse regularization based robust regression approaches,” *Signal Processing, IEEE Transactions on*, vol. 61, no. 5, pp. 1249–1257, 2013.
- [45] D. L. Donoho, M. Elad, and V. N. Temlyakov, “Stable recovery of sparse overcomplete representations in the presence of noise,” *Information Theory, IEEE Transactions on*, vol. 52, no. 1, pp. 6–18, 2006.